UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE INFORMÁTICA

Functional Coherence and Annotation Agreement Metrics for Enzyme Families

Hugo Paulo da Silva Bastos

DOUTORAMENTO EM INFORMÁTICA
ESPECIALIDADE BIOINFORMÁTICA

2015

UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE INFORMÁTICA

Functional Coherence and Annotation Agreement Metrics for Enzyme Families

Hugo Paulo da Silva Bastos

DOUTORAMENTO EM INFORMÁTICA
ESPECIALIDADE BIOINFORMÁTICA

Tese orientada pelo Prof. Doutor Francisco José Moreira Couto e pelo Prof. Doutor Luka Alexander Clarke, especialmente elaborada para a obtenção do grau de doutor em Informática, especialidade Bioinformática

2015

Resumo

Diversas metodologias são usadas para criar anotações em sequências, desde a curação manual por curadores especializados até vários procedimentos automáticos. A multitude de métodos de anotação existentes consequentemente gera heterogeneidade nas anotações em termos de cobertura e especificidade em espaços de sequências biológicas. Ao comparar grupos de sequências semelhantes (tais como famílias proteícas) esta heterogeneidade pode introduzir dificuldades quanto à interpretação da semelhança e coerência funcional nesses grupos. Uma maneira de mitigar essas dificuldades é a extensão da anotação dentro das famílias proteícas em análise. Esta tese postula que famílias proteícas podem ser usadas como bases de conhecimento para a sua própria extensão de anotação através do uso de análises de coerência funcional apropriadas. Portanto, uma *framework* modular para a análise de coerência funcional e extensão de anotação em famílias proteícas foi proposta. A *framework* incluí um módulo proposto para a análise de coerência funcional baseado em visualização de grafos, enriquecimento de termos e outras estatísticas. Neste trabalho o módulo foi implementado e disponibilizado como uma aplicação *web*, GRYFUN que pode ser acedida em http://xldb.di.fc.ul.pt/gryfun/. Adicionalmente, quatro métricas foram desenvolvidas para aferir aspectos distinctos da coerência e completude de anotação em famílias proteícas em conjunção com métricas já existentes. Portanto, o uso da *framework* completa por curadores, como uma estratégia de anotação semi-automática, é capaz de potenciar a extensão de anotação.

Palavras Chave: Anotação Funcional, Análise de Coerência de Anotação em Proteínas, Métricas de Anotação, Gene Ontology, Extensão de Anotações.

Abstract

A range of methodologies is used to create sequence annotations, from manual curation by specialized curators to several automatic procedures. The multitude of existing annotation methods consequently generates an annotation heterogeneity in terms of coverage and specificity across the biological sequence space. When comparing groups of similar sequences (such as protein families) this heterogeneity can introduce issues regarding the interpretation of the actual functional similarity and the overall functional coherence. A direct path to mitigate these issues is the annotation extension within the protein families under analysis. This thesis postulates that the protein families can be used as knowledgebases for their own annotation extension with the assistance of a proper functional coherence analysis. Therefore, a modular framework for functional coherence analysis and annotation extension in protein families was proposed. The framework includes a proposed module for functional coherence analysis that relies on graph visualization, term enrichment and other statistics. In this work it was implemented and made available as a publicly accessible web application, GRYFUN which can be accessed at http://xldb.di.fc.ul.pt/gryfun/. In addition, four metrics were developed to assess distinct aspects of the coherence and completeness in protein families in conjunction with additional existing metrics. Therefore the use of the complete proposed framework by curators can be regarded as a semi-automatic approach to annotation able to assist with protein annotation extension.

Keywords: Functional Annotation, Protein Annotation Coherence analysis, Annotation metrics, Gene Ontology, Annotation extension.

Resumo Alargado

Uma sequência biológica apenas adquire significado quando lhe é atribuído um contexto biológico apropriado. Esse contexto é tipicamente fornecido através de anotações funcionais, ou seja associações entre sequências e descriptores de funções exercidas pelas respectivas sequências biológicas. Diversas metodologias são usadas para criar anotações em sequências, desde anotações manuais de alta qualidade feitas por curadores especializados até vários procedimentos automáticos que tipicamente geram anotações mais genéricas. O compromisso actual entre curação manual e anotação automática dá-se através do uso de métodos semi-automáticos, onde procedimentos automáticos são empregues para propor anotações que requerem verificação por curadores especializados. A diversidade de diferentes métodos de anotação existentes consequentemente leva a uma heterogeneidade das anotações geradas relativamente à cobertura e especificidade das mesmas em agregados de sequências biológicas. Quando se compara grupos de sequências semelhantes (tais como famílias de proteínas) esta heterogeneidade pode introduzir dificuldades quanto à interpretação da semelhança e coerência funcional dentro desses grupos. Uma maneira de mitigar essas dificuldades é através da extensão da anotação dentro das famílias de proteínas em análise. Esta tese postula então que famílias de proteínas podem ser usadas como bases de conhecimento para a sua própria extensão de anotação através do uso de metodologias de análise de coerência funcional apropriadas.

Para esse efeito foi proposta uma *framework* modular de análise de coerência funcional e extensão de anotação em famílias de proteínas. Muitas das metodologias que levam à anotação funcional de proteínas (e outras biomoléculas) podem ser segregadas em dois meta-passos:

identificação de pares funcionais e *transferência de anotação*. Frequentemente as metodologias empregues para a construção de bases de dados especializadas de proteínas em domínios biológicos específicos sobrepõem-se com as metodologias usadas na tarefa *identificação de pares funcionais* em sistemas de anotação funcional. Além disso, essas bases de dados especializadas frequentemente são organizadas em famílias de proteínas. Famílias de proteínas são tipicamente grupos de sequências proteícas evolucionariamente relacionadas e que por consequinte é esperado nelas um certo grau de conservação funcional. Portanto a postulação de que famílias de proteínas podem ser usadas como base de conhecimento para a sua própria extensão funcional surge assente no conhecimento supracitado. É importante notar, que tal como para a maior parte dos outros métodos de anotação não existem anotações criadas *de novo*. O que acontece então é a extracção de conhecimento necessário para a inferência e extrapolação de termos de anotação já associados a algumas das proteínas da família para outras proteínas na mesma família ainda não anotadas com esses mesmos termos.

Assim sendo, pelos motivos acima descritos, podemos partir do princípio que uma família de proteínas terá sempre alta coerência funcional, mesmo que isso não seja imediatamente evidente pelo seu corrente estado de anotação. No entanto, essa coerência poderá ser estabelecida a diferentes níveis de especificidade funcional. Quando existe uma mistura de anotações de várias especificidades na mesma família poder-se-á então dizer que existe uma incompletude de anotação. Este tipo de anotação naturalmente pode causar dificuldades de interpretação funcional e várias métricas existem para aferir o grau de coerência funcional em grupos de proteínas. Porém muitas vezes essa coerência é medida a um nível de especificidade mais genérico e logo menos informativo. Adicionalmente, muitas proteínas são compostas por mais de um domínio e são multi-funcionais. Algumas dessas funções

poderão ser apenas assessórias e não relevantes para a caracterização funcional de uma determinada família de proteínas. Nestes casos esse tipo de funções assessórias, se contabilizadas em pé de igualdade com as restantes funções, poderão complicar mais ainda a aferição da coerência funcional de um determinado grupo de proteínas. Por estes motivos na metodologia proposta nesta tese, uma técnica de enriquecimento de termos (de anotação) foi aplicada. Tipicamente, esta técnica estatística é usada em casos de estudos diferenciais em *transcriptomics* onde se pretende determinar quais os genes expressos diferencialmente entre uma condição controlo e uma determinada condição em estudo. Paralelamente, neste trabalho é assumido que a criação de famílias (conjuntos) de proteínas dentro de determinadas colecções (bases de dados especializadas) gera enriquecimento de termos relevantes dentro das famílias. Para além disso, é assumido que os termos encontrados estatisticamente enriquecidos numa qualquer família são os caracterizadores funcionais dessa mesma família. Deste modo, esses termos são também os potenciais candidatos para extensão de anotação dentro dessa família.

A base de dados CAZy descreve os módulos (ou domínios funcionais) de famílias catalíticas estruturalmente relacionadas e de módulos de adesão a carbohidratos de enzimas que degradam, modificam ou criam ligações glicosídicas. A manutenção desta base de dados é feita por uma pequena equipa de curadores que usam métodos semi-automáticos para a manterem actualizada. Logo, esta base de dados oferece as propriedades ideais para servir como caso de estudo e desenvolvimento de métodos para a aferição de coerência funcional em grupos (famílias) de proteínas e portanto foi usada aqui como caso de estudo.

A aferição inicial do espaço de anotação de famílias do CAZy demonstrou a existência de incompletudes de anotação mas também oportunidades de extensão da mesma. A partir desse estudo uma *framework* modular para a análise de coerência funcional e extensão de

anotação em famílias proteícas foi proposta. A *framework* incluí um módulo para a análise de coerência funcional baseado em visualização de grafos, enriquecimento de termos e outras estatísticas. Por consequinte, este módulo foi implementado e disponibilizado como uma aplicação *web*, GRYFUN que pode ser acedida em http://xldb.di.fc.ul.pt/gryfun/. Adicionalmente o seu código fonte foi disponibilizado como *open source* sob uma licença MIT e depositado num servidor GIT em https://bitbucket.org/hpbastos/gryfunserver.git. O funcionamento desta aplicacação consiste em criar colecções de proteínas (que funcionam como fundo estatístico) e organizadas em conjuntos de estudo (de modo a replicar famílias de proteínas). Para cada conjunto de proteínas é então possível gerar o grafo de anotação (com os termos de qualquer uma das ontologias da *Gene Ontology*). Os grafos são criados por forma a evidenciarem o fluxo de anotação através da espessura das arestas. Para complementar cada grafo gerado, é apresentada uma tabela com os resultados do enriquecimento dos termos com o *p-value* respectivo para cada termo, a contagem de anotações e uma métrica baseada no conteúdo de informação. Em termos de interacção a funcionalidade mais inovadora no GRYFUN é a capacidade de iterativamente gerar sub-grafos a partir de um grafo inicial permitindo focar a atenção a sub-grupos funcionalmente mais próximos.

Embora que ainda não implementados na aplicação GRYFUN, quatro outras métricas adicionais foram desenvolvidas para aferir aspectos distinctos da coerência e completude de anotação em famílias de proteínas em conjunção com métricas já existentes. Duas das métricas, *IC-completeness* and *Leaf-completeness* abordam de uma forma *naïve* o problema de completitude, e são reforçadas pelas técnicas de visualização implementadas na aplicação GRYFUN. As outras duas métricas, *mUI* e *mGIC* derivam de uma combinação individual híbrida entre duas métricas de semelhança semântica e técnicas de enriquecimento de termos. Estas duas métricas permitem o focar na coerência

funcional local dentro de conjuntos (ou famílias) de proteínas.

O módulo para extensão de anotação em famílias de proteínas depende então dos resultados produzidos pelo módulo anterior. Portanto, os sub-grupos identificados dentro de cada família são usados para criar primeiro alinhamentos múltiplos de sequências (através do programa MAFFT) que depois são convertidos em perfis de Modelos Escondidos de Markov (através do programa HMMER). Estes perfis são depois usados para tentar classificar proteínas sub-anotadas dentro de uma família. Os ensaios executados demonstraram uma precisão total e um *recall* variável, parcialmente dependente do número de sequências usadas para criar o respectivo alinhamento múltiplo de sequências.

Logo, ficou demonstrado que o uso, por curadores, da *framework* completa que é proposta aqui como uma estratégia de anotação semi-automática, tem a capacidade de potenciar a extensão de anotação em famílias de proteínas.

Palavras Chave: Anotação Funcional, Análise de Coerência de Anotação em Proteínas, Métricas de Anotação, Gene Ontology, Extensão de Anotações.

Acknowledgements

First, I would like to thank my advisors Francisco Couto and Luka Clarke for their constant support, ready availability and invaluable encouragement throughout the course of this thesis.

I would also like to thank Pedro Coutinho that was my co-advisor during the first part of my thesis and made me feel welcomed in each of the visits to his lab.

Furthermore, I am also thankful to André Falcão, that with his overwhelming enthusiasm introduced and led my first steps into scientific research.

I am also grateful to *Fundação para a Ciência e a Tecnologia* that through PhD Grant ref. SFRH/BD/48035/2008 financially enabled me to pursue the research work necessary for this thesis.

I am immensely thankful to all the people that shared this journey with me, most of them starting as colleagues, then collaborators and all ending up as good friends. Among these I highlight Cátia Pesquita, Daniel Faria, Tiago Grego, João Leitão, Cátia Machado, Bruno Tavares, Luís Filipe, Juliana Duque, Ana Teixeira, João Ferreira and Sílvia Mendonça. I am also thankful to a great number of friends that will remain unnamed for obvious practical reasons.

A special thanks for an odd bunch of friends Joseph Long, Nathaniel Morton, Neil Skrypuch, Travis Cavanaugh for their mostly random but valued support.

And last, but not least my family. I am endlessly thankful to my parents, David and Anabela, for the obvious and also less obvious reasons. Also I thank Sónia, I couldn't wish for a better little sister. And a very special loving thanks to Greyce, my wonderful fiancée, that stood by me and always believed me, lending me strength when I needed it the most.

"The important thing in science is not so much to obtain new facts as to discover new ways of thinking about them."

Sir William Bragg

(1862 - 1942)

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The fast pace at which genomes are being fully sequenced is greatly increasing the amount of available gene product sequences. These sequences are strings of characters that represent the biochemical material resulting from the expression of a gene. The resulting material can be either protein or RNA. Proteins play an important part in organisms and participate in practically every process occurring within living cells. Their functions range from structural or mechanical support to catalysis of vital metabolic biochemical reactions. Proteins are usually sorted into evolutionarily-related groups called *protein families*. Within a family each protein is homologous to all the other proteins, i.e., it descends from a common ancestor and typically retains significant sequence similarity which in turn can translate into similar three-dimensional structures and functions. While sequence similarity alone is not sufficient to conclude protein homology, it nevertheless provides a reasonable cornerstone for many sequence alignment methods making them useful in determining homology relationships. Hence, many algorithms were devised to be able to cluster protein sequences into families of homologue sequences. However, the definition of *protein family* still carries some usage ambiguity and is context-dependent. Thus, according to different researchers a *protein family* can either indicate small groups of proteins with nearly identical sequence, structure and function up to larger groups of distantly related proteins retaining only enough sequence similarity to be detected as distant homologues. For the purposes of this work *protein families* are defined as sets of proteins, more specifically enzymes (or their functional domains) that share considerable sequence and

functional similarity. Furthermore, enzymes within a protein family that have even more closely related functions can be considered as constituting sub-families (Stam *et al.*, 2006). That definition is also adopted and used throughout this manuscript.

However, grouping protein sequences is not sufficient to provide full biological context. Generally, it is necessary either to experimentally determine or predict their function. The specification of protein function is very broad and can range from descriptors on participation in biological processes such as responses to oxidative stress up to specific descriptors on the catalysis of biochemical reactions. Furthermore, proteins can have one or several biochemical functions and also participate in numerous biological processes. Optimally, thorough and dedicated chemical characterizations would be used to determine the functions of proteins but this approach is expensive and time consuming. A more commonplace approach is the use of any of the several function prediction methods relying on techniques ranging from sequence homology detection to text mining of the scientific literature. Most of these methods make use of computational procedures that enable handling the barrage of biological data currently being made available. Given a supporting evidence, predictions can be obtained (with varying degrees of confidence) and thus functional descriptors can be assigned to the proteins. In the Biological Sciences this process is referred to as *protein annotation* whereas a protein functional annotation is actually the *protein sequence - functional descriptor* pair. Over the last decades, the functional descriptors used for gene product functional annotation have shifted from the initial free text annotation model to annotation using terms from controlled and structured vocabularies. The standardization of annotation terms is particularly useful for the increasing usage of automated annotation methods. This in turn leads to an ever increasing availability of protein annotations. Currently, the Gene Ontology (GO) Consortium (Gene Ontology Consortium, 2000) sets the standard in the community when it comes to providing a controlled vocabulary of terms to describe genes and attributes of gene products of any organism. The increasing popularity of GO terms for protein annotation also led to the development of several associated semantic similarity based metrics. Lord *et al.* (2003) were

the first to apply semantic similarity based on GO in order to compare gene products based on their functional information. They adapted and tested three measures: Resnik's (Resnik, 1995), Lin's (Lin, 1998), and Jiang and Conrath's (Jiang & Conrath, 1997) that were originally developed for the WordNet (Miller, 1995) taxonomy, a lexical database for the English language. Within the GO annotation context (and also in this manuscript), semantic similarity can then be defined as the closeness in meaning between two terms or two sets of terms annotating two proteins. Hence, the semantic similarity between two proteins annotated with GO terms can also be referred as *functional similarity*. Over the last few years several semantic-based measures have been developed and used to assess functional closeness or coherence within protein sets. However, caution must be exerted during the interpretation of protein set results for functional coherence assessments. Differences in annotation specificity, a form of annotation incompleteness, can occur for any set of functionally related proteins. That can lead to semantic similarity measures reporting low similarity values, however they generally cannot be interpreted as differences in actual protein function. There is also an inherent research bias towards the more intensively studied model organisms. As such, different methods are employed towards protein annotation and thus different functional terms of varying specificities are assigned to potential homologue proteins according to the confidence thresholds used. Recently, Nehrt *et al.* (2011) proposed and applied a metric for "functional similarity" based on GO annotations. Applying their metric on two genomes (human and mouse) the authors concluded, quite unexpectedly, that on average groups of genes commonly linked to function innovation (diverging functions) showed more semantic similarity than groups where function conservation is typically expected. Thomas *et al.* (2012) challenged these conclusions and although confirming the observed annotation agreement they also showed that the difference was derived from incomplete and complementary annotations. Thus, the development of functional similarity measures able to gauge the state of annotation incompleteness (due to annotations of varying specificities) within a set of functionally related proteins becomes much required. These measures would allow to assess the true value of protein (or other gene products) annotation similarity comparisons.

1.1 Objectives

The main goal of this work is to mitigate issues stemming from the previously mentioned annotation incompleteness resulting from the heterogeneity in annotation specificity within sets (families) of functionally related proteins. In order to more reliably and accurately use protein functional annotations as a measure of functional similarity this type of annotation incompleteness has to be taken into account. Otherwise, semantic similarity measures may only reflect annotation disagreement not indicative of actual functional differences. On the other hand, the optimal strategy to mitigate issues stemming from annotation incompleteness is to reduce that incompleteness to the further extension possible.

Hypothesis: It is possible to extend functional annotations in protein families with the assistance of adequate functional coherence analysis considering that families are expertly collected knowledgebases.

Therefore, this thesis proposes to advance the state of the art with contributions on the use of semantic similarity and enrichment techniques applied to the extension of protein annotation of expertly-created protein collections, such as the ones comprised of evolutionarily-related groups, or families.

1.2 Methodology

A modular framework for the extension of annotation within proteins families and the measuring of their respective functional coherence was conceived. Instantiation of modules for that framework was done and thus modules were developed and used to test and validate the hypothesis.

The CAZy (Cantarel *et al.*, 2009) database describes families of structurally-related catalytic and carbohydrate-binding modules of enzymes that degrade, modify, or create glycosidic bonds. The statistical characterization of its annotation corpus provided insight both on the state of annotation of an expertly created protein family database but also on the use of semantic similarity in order to measure intra-family functional relatedness.

The statistical approach of the term enrichment was merged with semantic similarity, and GRYFUN, a web-based application was developed combining these merged techniques with graph visualization. Additional functional coherence metrics, mUI and mGIC were developed based on the combination of term enrichment merged with semantic similarity. The behaviour of these metrics was tested alongside others, for baseline and comparison, against CAZy families that were increasingly transformed into random sets.

An annotation extension module was developed relying on the MAFFT software package (Katoh & Toh, 2008) for the creation of automatic multiple sequence alignments (MSA) and the HMMER software (Finn *et al.*, 2011) to build profile hidden Markov models out of the MSA and use them to propose the extension of annotation terms to sub-annotated proteins in families. This module was validated through repeated random sub-sampling validation and performance evaluated through the standard evaluation metrics: precision, recall and F-measure.

As a final task the proposed framework was evaluated using selected CAZy and MEROPS (Rawlings *et al.*, 2012) (release 9.9) database families which were submitted through the whole pipeline of instantiated modules followed by a verification of the obtained results against domain knowledge.

1.3 Contributions

The endeavours undertaken during the process of validating the hypothesis have led to contributions in the areas of semantic similarity application, functional coherence measuring and protein annotation. The contributions can be summarily described as follows:

- **Semantic similarity** Two novel hybrid sub-local similarity metrics, mUI and mGIC, were developed for the purpose of tracking high-similarity subsets of proteins within protein families.

- **Functional coherence analysis** GRYFUN, a web-based application merging graph visualization, term enrichment, statistical-based and semantic-based metrics was developed in order to perform functional annotation coherence analysis in protein sets (families).

- **Functional annotation extension** The direct way to deal with the heterogeneous annotation specificity is to homogenize the annotation specificity within a family. A methodology using Hidden Markov Model (HMM) profiles was developed to harness the results of the functional analysis developed in this thesis and allow the extension of annotation in protein families.

- **Framework for annotation coherence and extension in proteins families** Finally, each of the previously described contributions was implemented as a module capable of being integrated in the proposed modular framework for the functional coherence assessment and annotation extension within protein families (Bastos *et al.*, 2013, 2015).

1.4 Overview

The remainder of this thesis is organized according to the following structure:

Foundations The theoretical basis and contextual work that motivates this thesis is presented. Chapter 2 introduces the basic notions of Molecular Biology needed to understand the underlying biological motivations and implications. Also in this chapter, public protein database management, functional annotation schemas and semantic similarity metrics are introduced and described in order to later make clear the proposed approaches of this thesis. Chapter 3 presents the state-of-the-art in protein annotation which drives and motivates the creation of a functional annotation coherence measuring methodology. The state-of-the-art in semantic similarity (and otherwise) metrics applied to the functional coherence

measurement is also presented and discussed.

Methodology Regarding Chapter 4 the exploratory analysis of a protein database annotation space is described, followed by the study of a set of semantic similarity-based metrics applied to protein family (set) functional coherence measurement. Within this chapter a framework for protein extension in protein families guided by functional coherence measurement is proposed. In Chapter 5 the web-based application GRYFUN, a annotation graph visualization and term enrichment tool designed with particular focus on annotation coherence analysis and annotation extension for protein families is presented and described.

Conclusions Chapter 6 demonstrates the feasibility of the complete pipeline of modules in the proposed framework. Finally, Chapter 7 presents the overall summary of the contributions of this thesis, future directions and some closing remarks.

Chapter 2

Basic Concepts

Understanding the interactions between the various systems of a cell, such as the interactions between DNA, RNA and protein biosynthesis along with learning how these interactions are regulated is the main purpose of Molecular Biology. Bioinformatics appears as an interdisciplinary research area at the interface between Biology, Biochemistry, Biophysics, Statistics, Mathematics, and Informatics. Bioinformatics proposes to tackle the challenges presented by Molecular Biology. Several research fields are encompassed by both Molecular Biology and Bioinformatics, however this Chapter only covers the topics, needed for the understanding of the work presented in this document. Such topics will be briefly covered and some biological facts may not hold true for all biological systems, however exceptions outside the scope of this work will not be indicated for the sake of readability.

This Chapter is organized as follows. Section 2.1 reviews the basic concepts of Molecular Biology and the information it generates. Section 2.2 explains how public databases manage, maintain and make available this information, especially protein information. Section 2.3 describes the popular sources for functional schemas commonly used in protein annotation. Section 2.4 elaborates on GO term annotations in particular, while Section 2.5 describes the use of these same annotations in semantic similarity metrics. Finally, Section 2.6 summarily describes GO term enrichment analysis.

2.1 Molecular Biology

Living organisms are dependent on complex and essential information storage and processing. That information is typically stored within genes inside the organism's cells and used for cell maintenance and genetic trait transmission to the offspring. Effective storage, expression, and reproduction of the genetic information defines individual species, distinguishes them from one another, and assures their continuity over successive generations (Nelson & Cox, 2004). Currently, a gene can be defined as a locatable region of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions, and or other functional sequence regions (Pearson, 2006). Biochemically, genes are composed by double strands of deoxyribonucleic acid, DNA. Each DNA molecule is a linear polymer of four different monomeric subunits, desoxyribonucleotides, arranged in a precise linear sequence. Desoxyribonucleotides are constituted by a common deoxyribose sugar and a phosphate group and differentiated by a nitrogenous base that can be either be adenine, thymine, guanine, and cytosine. It is this linear sequence that encodes the genetic information. Two of these polymeric strands are twisted about each other to form the DNA double helix, in which each deoxyribonucleotide in one strand pairs specifically with a complementary deoxyribonucleotide in the opposite strand (Nelson & Cox, 2004). Under the commonly used IUPAC system, nucleic acids are represented by the first letters of their chemical names: Guanine, Cytosine, Adenosine, and Thymine (IUPAC-IUB Comm. on Biochem. Nomenclature (CBN), 1970). Hence a DNA strand is usually represented as a string of the letters G, C, A and T. Figure 2.1 shows the *Homo sapiens* mRNA for prepro cortistatin-like peptide, complete coding sequence (CDS).

DNA is commonly found in the nucleus of (eukaryote) cells as organized structures, i.e. chromosomes. Chromosomes are pieces of coiled DNA, each containing many genes, regulatory elements and other nucleotide sequences. Chromosomes are stabilized by DNA-bound proteins, called histones, which serve to package and control the functions of DNA. Figure 2.2 illustrates the generic structure of chromosomes and their DNA packing.

```
ACAAGATGCC ATTGTCCCCC GGCCTCCTGC TGCTGCTGCT CTCCGGGGCC ACGGCCACCG
CTGCCCTGCC CCTGGAGGGT GGCCCCACCG GCCGAGACAG CGAGCATATG CAGGAAGCGG
CAGGAATAAG GAAAAGCAGC CTCCTGACTT TCCTCGCTTG GTGGTTTGAG TGGACCTCCC
AGGCCAGTGC CGGGCCCCTC ATAGGAGAGG AAGCTCGGGA GGTGGCCAGG CGGCAGGAAG
GCGCACCCCC CCAGCAATCC GCGCGCCGGG ACAGAATGCC CTGCAGGAAC TTCTTCTGGA
AGACCTTCTC CTCCTGCAAA TAAAACCTCA CCCATGAATG CTCACGCAAG TTTAATTACA
GACCTGAA
```

Figure 2.1: *Homo sapiens mRNA for prepro cortistatin-like peptide, complete CDS*

The continued existence of a biological species requires its genetic information to be maintained in a stable form, expressed accurately in the form of gene products, and reproduced with a minimum of errors (Nelson & Cox, 2004). Identical copies of DNA must be produced in order to transmit the genetic information to new cells and progeny which is achieved through the DNA replication mechanism. In this process each strand of the original double-stranded DNA molecule serves as template for the reproduction of the complementary strand. Regardless of type, most cells in a multi-cellular organism contain identical DNA. However, depending on cell cycles, environments and external signals different cells can have differential sets of genes active. Several genes are typically protein encoding genes. The transcription and translation mechanisms are responsible for the protein synthesis. The (almost) uni-directional flow of information necessary to produce the proteins is known as the central dogma of Molecular Biology (Crick, 1958). The replication, transcription and translation mechanisms control this process as illustrated by Figure 2.3. Transcription occurs within the nucleus where with the segment of DNA encoding a gene is first transcribed to a polymer of messenger ribonucleic acid (mRNA). This mRNA molecule is a temporary intermediary that is transported to the cytoplasm where it is translated into proteins by specialized cellular components, the ribosomes. Figure 2.4 illustrates the processes of DNA transcription and subsequent mRNA translation into proteins.

Proteins are organic compounds, polymers of amino acids (polypeptides) arranged in a linear chain joined together by the peptide bonds between the carboxyl and amino groups of adjacent amino acid residues. In general, the genetic code specifies 20 standard amino acids. Just like for the nucleotides, each one of

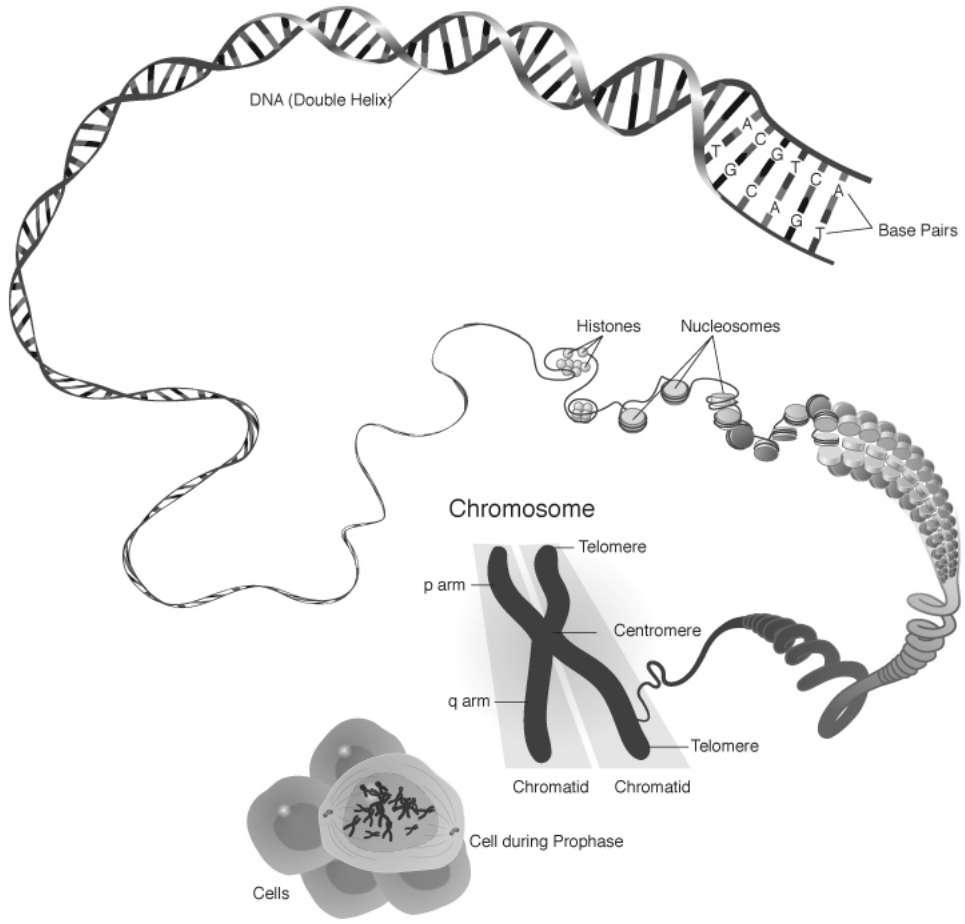


Figure 2.2: *Double anti-parallel chains of DNA joined by hydrogen bonds twist into a double helix and pack around histones forming one of the cell's nuclear chromosomes.*

the natural amino acid residues can also be denoted by a code of single Roman alphabet characters (along with an existing abbreviation form of a 3 letter code). On mRNA molecules each three sequential nucleotides (within a proper reading frame) is called a codon. During the translation process mRNA goes through the ribosome and each codon is paired up with its anti-codon. Each anti-codon is composed by three sequential nucleotides complementary to those of a codon and which are located on transfer RNA (tRNA) molecules. These RNA molecules carry specific amino acids that are then added to the protein chain being synthesized. Hence, in a protein sequence each amino acid is directly encoded by three nucleotides on a DNA sequence.

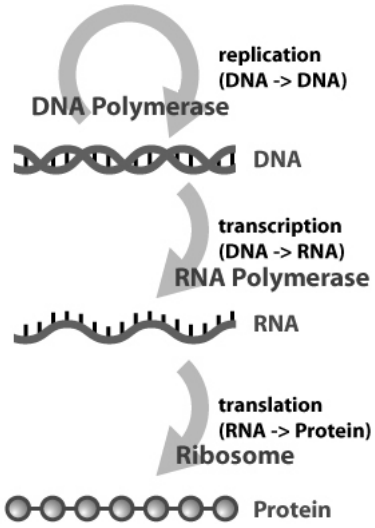


Figure 2.3: *The central dogma of Molecular Biology.*

After synthesis, protein chains normally fold into 3-dimensional structures. The result corresponds to native conformations, that are ultimately determined by the order and chemical properties of the amino acid residues that form the chain or chains. Although many proteins can fold without any help, other proteins need the assistance of special proteins, named chaperones to achieve native conformation. Protein structure can be divided into four different tiers, which are illustrated in Figure 2.5.

- Primary structure: the amino acid sequence;
- Secondary structure: the common repeating local structures, such as the alpha helices, beta sheets or loops. Since they are local, different kinds of structure can be present in the same protein;
- Tertiary structure: the overall conformation of a single protein chain. The terms "tertiary structure" and "fold" are often used as synonyms, although the latter more often describes the mode by which the peptide folds itself.
- Quaternary structure: some proteins have a final structure formed by several protein chains, or protein subunits in this context, which may function as a single protein complex.

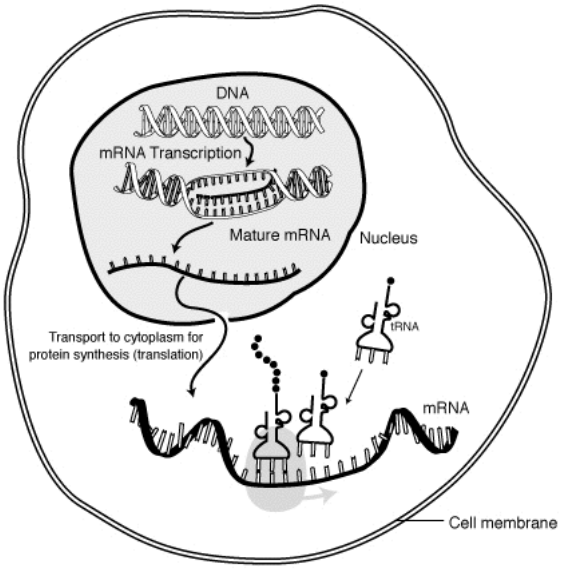


Figure 2.4: *A segment of protein-encoding DNA is transcribed at the nucleus into a messenger RNA (mRNA) and transported to the cell's cytoplasm. The mRNA is then translated into a protein in the ribosomes.*

Proteins are essential components of organisms and participate in practically every process occurring within living cells. Numerous proteins are enzymes that catalyse biochemical reactions and are vital to the metabolism. The processes in which they participate range from cell signalling, immune responses, cell adhesion to the cell cycle. Some proteins also have structural or mechanical functions. A crucial step following full genome sequencing of an organism is typically its functional annotation. The identification of gene and protein activities can lead to a better understanding of how living systems work (Nowak, 1995). Protein function can be considered at different levels. From a more specific level each protein has an elementary biochemical function like a catalytic or binding activity. At a upper functional level biological roles are performed by groups of proteins working together towards a common purpose, for instance, the signal transduction pathways. Nonetheless, the identification of gene and proteins activities is a non-trivial task. Biological systems have numerous genes and proteins interacting in complex ways to regulate one another and adjusting themselves according to environment changes, such as molecular signals or physiological conditions. However, when

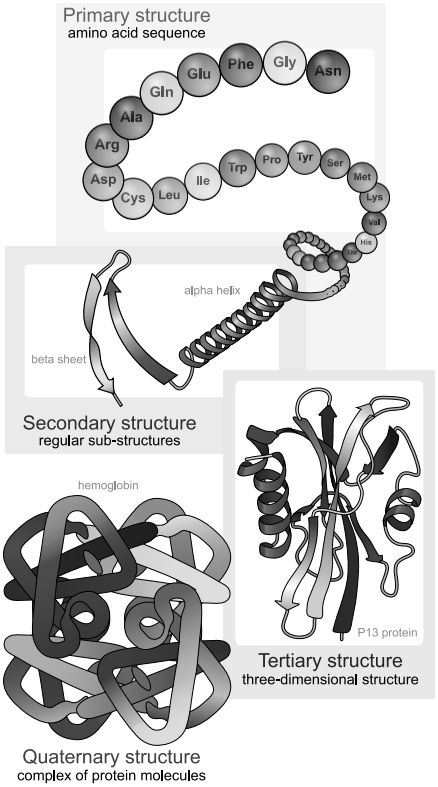


Figure 2.5: *Proteins fold into 3D structures. The structure of proteins can be divided into four aspects here depicted.*

proper evidence is available functional descriptors based on that evidence can be associated to the database identifiers of the respective involved proteins (or other gene products). This evidence-based *protein - functional descriptor association* defines the concept of *functional annotation* used throughout this document.

2.2 Protein databases

Biological sciences have been changed by the development of high-throughput technologies for biological research. The current deluge of information generated can no longer be contained just by the traditional biological literature. As of September 2011 there are over 2900 completely sequenced and published genomes

and more than 11,400 genome projects described in GOLD [1] (Pagani *et al.*, 2012). Figure 2.6 shows the progress in the number of completed genome sequencing projects over the years.

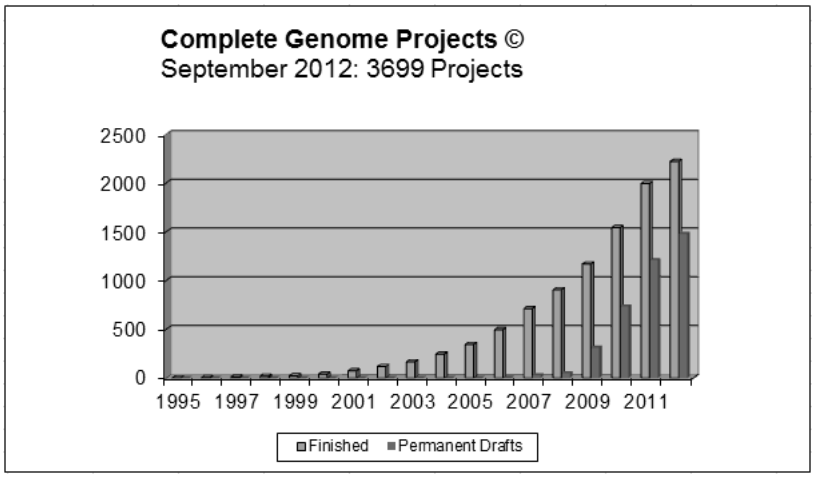


Figure 2.6: *Timeline of completely sequenced genome projects*

Along with the growth in biological data output also came the corresponding increase in the number and size of public databases designed to maintain that data. Currently, a myriad of databases is available on the Web. However, most of these databases are based on data initially stored at a few central databases. These *primary* databases have the purpose of mainly storing sequence and structural information of genes and proteins, but also can contain functional information. Traditionally, the main *primary* nucleotide databases are:

DNA Data Bank of Japan (DDBJ) : Asia's sole nucleotide sequence data bank, and which is part of the consortium of databases that collects nucleotide sequences from the research community.

EMBL Nucleotide Sequence Database (EMBL-Bank) : Europe's primary nucleotide sequence resource. Main sources for DNA and RNA sequences are direct submissions from individual researchers, genome sequencing projects and patent applications (Cochrane *et al.*, 2009).

[1] http://www.genomesonline.org/

GenBank : the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences for more than 300 000 organisms named at the genus level or lower, obtained primarily through submissions from individual laboratories and batch submissions from large-scale sequencing projects (Benson *et al.*, 2008)

The data collected by these data banks is exchanged on a daily basis, hence each one of them share virtually the same data at any given time. This virtually unified data bank is called the International Nucleotide Sequence Databases (INSD) and has been collaboratively developed and maintained for over 18 years by DDBJ, EMBL, and GenBank (Nakamura *et al.*, 2013).

The Universal Protein Resource(UniProt) : a central resource on protein sequences and functional annotation created with the union of the Swiss-Prot, TrEMBL, and PIR protein database activities (The UniProt Consortium, 2013). It has four major database components, each addressing a key need in protein bioinformatics. Of these components, the UniProt Knowledgebase (UniProtKB) is the preeminent storehouse of protein annotation. It comprises the manually annotated UniProtKB/Swiss-Prot section and the automatically annotated UniProtKB/TrEMBL section. The extensive cross-references, functional and feature annotations, and literature-based evidence attribution enable scientists to analyze proteins and query across databases.

The Protein Data Bank (PDB) : a repository of information about the 3D structures of biological macromolecules, including proteins and nucleic acids (Berman *et al.*, 2007). This database presents a high level of redundancy as a single protein may be present in different crystal forms, different mutants and different complexes with other proteins and ligands. As of July 1st, 2012 there were over 76,500 protein/peptide structures hosted by PDB (Rose *et al.*, 2013).

Further biological databases can be found in the Nucleic Acids Research Database Categories List. [1]

The information present on *primary*, or general sequence databases is usually replicated and distributed over a multitude of *secondary* databases. However, these databases usually have more specific purposes and aim at organizing, integrating and classifying all, or just part of the information stored in *primary* databases. Thus, each *secondary* database is typically focused on particular specific subject ranging from general protein properties to databases of individual protein families. Throughout that range some of these databases conveniently aggregate sequence data such as localization, motifs, active sites, domain information and protein classification which can be particularly useful for functional inference purposes. Below, some common examples of *secondary* databases that typically provide resources for several of the existing biological sequence functional annotation systems are briefly described.

Blocks : a database of multiple alignments, which represents conserved protein regions, that is, identical or very similar sequences. Blocks are automatically constructed by looking for the most highly conserved ungapped segments in the protein families documented in InterPro and can help verify homology (Henikoff, 2000).

PRINTS : a compendium of protein family fingerprints, a fingerprint being a group of conserved motifs used to characterize a protein family. It is also specialized in provisional hierarchical classifications of protein superfamilies (Attwood, 2003).

PROSITE : a database consisting of documentation entries describing protein domains, families and functional sites as well as associated patterns and profiles to identify them. PROSITE is complemented by ProRule, a collection of rules based on profiles and patterns, which increases the discriminatory power of profiles and patterns by providing additional information about functionally and/or structurally critical amino acids (Sigrist *et al.*, 2010).

[1] http://www3.oup.co.uk/nar/database/c/

ProDom : a comprehensive set of protein domain families automatically generated from the UniProtKB (Bru *et al.*, 2005). This resource uses profiles in the form of position specific scoring matrices constructed using PSI-BLAST (Altschul *et al.*, 1997).

STRING : a database of known and predicted protein-protein interactions. The interactions include direct (physical) and indirect (functional) associations; they are derived from four sources: genomic context; high-throughput experiments; co-expression; and previous knowledge from databases and the scientific literature (Jensen *et al.*, 2009).

COG : Clusters of Orthologous Groups of proteins are delineated by comparing protein sequences encoded in complete genomes, representing major phylogenetic lineages. Each COG consists of individual proteins or groups of paralogues from at least 3 lineages and thus corresponds to an ancient conserved domain (Makarova *et al.*, 2007). COGs are derived from prokaryotes and KOGs from eukaryotes.

SMART : (Simple Modular Architecture Research Tool) provides high quality, manually curated Hidden Markov models (HMMs) and alignments of protein domain families. It facilitates identification and annotation of genetically mobile domains and the analysis of domain architectures. These domains are extensively annotated with respect to phyletic distributions, functional class, tertiary structures and functionally important residues (Letunic *et al.*, 2012).

Pfam : a large collection of protein families, each represented by multiple sequence alignments and HMMs. This resource has two components, Pfam A, where the entries are high quality, manually curated families and Pfam B consisting of automatically generated entries (Finn *et al.*, 2010).

TIGRFAMs : collection of curated multiple sequence alignments, HMMs and associated information designed to support automated annotation of (mostly prokaryotic) proteins. Cutoff scores and membership in the seed alignment

are chosen so that the HMMs can classify proteins according to their specific molecular functions (Haft *et al.*, 2013).

CAZy : a database that describes the families of structurally-related catalytic and carbohydrate-binding modules (or functional domains) of enzymes that degrade, modify, or create glycosidic bonds (Cantarel *et al.*, 2009). Its maintenance is done by a small team of curators that uses semi-automatic methods to keep it up-to-date. Even with part of the procedure being automatic there is still a large workload of manual curation that has to be performed by the specialized curators. Recently, the CAZy database has shifted from a schema where function was attributed to the complete enzyme sequence to a schema where function may be assigned just to the segment of the sequence involved in each function, the *functional module*. So far the CAZy families have been functionally annotated with Enzyme Commission (EC) numbers (Webb *et al.*, 1992). The EC number is a numerical classification for enzymes, based on the reactions they catalyze. Lately, the CAZy curators started to develop and use a new vocabulary of functional annotation terms organized into a tree structure to annotate proteins. This new structure allows not only to cope with enzymatic activities that have not yet been described by the EC, but also better organizes (hierarchically) the activities that in several families are just a subset of already described activities. Furthermore, this structure can also describe non-enzymatic functions or activities, such as the ability to bind carbohydrates displayed by individual Carbohydrate-Binding Modules (CBM). Functions can now be assigned not only to the protein but within the protein may be assigned to individual functional modules that make up the CAZy families. The module-centric organization schema of the database can be complemented in a way that functions, enzymatic or not, may be directly assigned to a specific segment of a sequence. The protein (enzyme) families found on the CAZy (`www.cazy.org`) database will be used as the main case-study in this work.

Commonly, the *primary* databases link their data to complementary information found in *secondary* databases. These databases, since they are usually specialized often contain additional experimental data not present in the *primary*

databases. The additional data brought forward to the *secondary* databases often proves essential for understanding the biological roles of genes and proteins. That additional data is commonly found and extracted from the biological literature through careful analysis by expert curators. However, the traditional functional characterization of genes and proteins cannot keep up with the large output of sequencing projects. Hence, automatic tools have been developed and used in order to extrapolate functional annotations from similar already functionally characterized sequences. Despite enhancing the rate of sequence annotation these tools also produced a significant number of misannotations that are now present in the databases (Devos & Valencia, 2001). One typical problem capable of leading to misannotations, is the undistinguished use between extrapolated and curated annotations by some of these tools to extrapolate new annotations without provenience information. Furthermore, when full sequences rather than functional domains are used the errors of functional transfer based on the similarity to adjacent modules also increases. Additionally, the lack of linking of experimental evidence to functional characterization makes error detecting more difficult. The detection and removal of errors can also be frequently impaired by the lack of standard nomenclature across biological databases rendering cross-checking ineffective.

2.3 Functional schemas

Functional annotations of biomolecules are evidence-based associations of biological activity or function descriptors with the database identifiers of their respective biomolecules. Sometimes these annotations are stored in biological databases as statements that are very domain specific and context dependent. Early protein annotations were just free-text functional descriptors being assigned by research teams performing experimental protein functional characterization. However, without a controlled vocabulary this kind of annotation relied too much on the subjectivity of individual researchers. Since then several classification schemas have been developed to control the annotation of biological entities. Below, three popular initiatives that deliver commonly used classification schemas for proteins (and other gene products) are briefly presented.

IntEnz is the Integrated relational Enzyme database and official version of the Enzyme Nomenclature (Fleischmann *et al.*, 2004). The Enzyme Nomenclature comprises recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB) on the nomenclature and classification of enzyme-catalysed reactions using EC (Enzyme Commission) numbers (Webb *et al.*, 1992). Therefore, strictly speaking EC numbers specify enzyme-catalyzed reactions and not the enzymes themselves, different enzymes catalyzing the same reaction receive the same EC. IntEnz, currently also integrates ENZYME, a repository of information relative to this nomenclature of enzymes (Bairoch, 2000). IntEnz is supported by NC-IUBMB and contains enzyme data curated and approved by this committee.

FunCat a functional catalogue initially developed for the *Saccharomyces cerevisiae* genome project at the Munich Information Center for Protein Sequences (MIPS) but since expanded to other organisms. It comprises of 28 main functional categories (or branches) that cover general fields like cellular transport, metabolism and cellular communication/signal transduction. Each branch exhibits a hierarchical, tree like structure with up to six levels of increasing specificity (Ruepp *et al.*, 2004).

KEGG The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a database of biological systems that integrates genomic, chemical and systemic functional information. Molecular building blocks are distinguished between genetic building blocks (KEGG GENES) and chemical building blocks (KEGG LIGAND), while the systemic information is represented as molecular wiring diagrams (KEGG PATHWAY) and hierarchies and relationships among biological objects (KEGG BRITE) (Kanehisa *et al.*, 2006).

GO The Gene Ontology (GO) project provides a structured and controlled vocabulary to describe genes and gene products in terms of their associated biological processes, cellular components and molecular functions (Gene Ontology Consortium, 2000). The GO project due to its broad scope and

wide applicability currently represents the most popular ontologies for describing gene and protein biological roles. Each of the three orthogonal aspects that compose GO describe gene product phenomena at different levels. Proteins typically have elementary molecular functions that are independent of the surrounding environment, such as catalytic or binding activities, these are described by the *molecular function* aspect. On the other hand, sets of proteins interacting and involved in cellular processes, such as metabolism or signal transduction are described by the *biological process* aspect. Proteins can perform their functions in several cellular localizations, such as the Golgi complex or the ribosome which are described by the *cellular component* aspect.

2.4 GO annotations

The GO project aims at providing generic consistent descriptions for the molecular phenomena in which the gene products are involved. The genesis of GO consisted of adding generic terms connected by simple relationships to the GO ontologies. That process enabled a broad coverage over the Molecular Biology domain fields. Nonetheless, the lack of more specific terms pertaining to some domains of Molecular Biology may hamper the usefulness of GO. However, the GO ontologies are dynamic regarding content and are being kept updated by the GO Consortium members, with currently over 40,000 terms. Furthermore, as the different research communities understand the importance of adding their domain knowledge to GO, it will acquire even more specific terms and relationships thus overcoming this limitation. The three biological aspects, biological processes, molecular functions and cellular components encompassed by GO are each represented by an individual Directed Acyclic Graph (DAG) where each node represents a term and the edges represent a relationship between those terms. Each term is identified by an alphanumeric code (e.g., GO:0003674) and its textual descriptors, including its name, definition, and synonyms when they exist. The currently available relationships between terms can be of three types: *is-a*, *part-of* and *regulates*. Figure 2.7 shows a sub-graph of the GO *biologi-*

cal_process aspect. Only *is-a* relationships are depicted in this example.

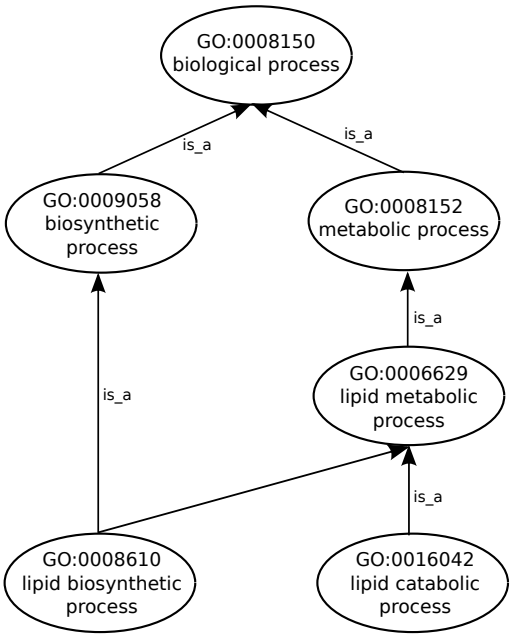


Figure 2.7: *Sub-graph of GO biological_process aspect.*

Proteins (or other gene products) are not actually part of GO, which includes only the terms that describe their functions or activities. However, the GO Consortium, through the Gene Ontology Annotation (GOA) project (Dimmer *et al.*, 2012) provides high-quality electronic and manual annotations, that is, associations of GO terms to UniProtKB entries. Annotations created by the GOA project are collated with annotations from external databases to provide an extensive, publicly available GO annotation resource. Currently supplying over 100 million annotations to 11 million proteins in more than 360,000 taxa, GOA remains the largest and most comprehensive open-source contributor to the GO Consortium project. Each protein can be annotated with as many GO terms required to fully describe its function(s). Considering Figure 2.7 any protein annotated with *lipid catabolic process* according to GO's true path rule ("the pathway from a child term all the way up to its top-level parent(s) must be true") then can also inherently be automatically and correctly annotated to all its parents (*lipid metabolic process*, *metabolic process* and *biological process*) up until

the root term of the respective DAG. Furthermore, each annotation associating a GO term to a protein is also attributed an evidence code, which is a 3-letter acronym identifying the type of evidence supporting the annotation (e.g. IPI: Inferred from Physical Interaction is assigned to annotations supported by this kind of experiment). Annotations are sometimes discriminated based on their evidence codes into two main types: manual annotations and electronic annotations. Manual annotations correspond to annotations manually made by expert curators, whereas electronic annotations are inferred by automatic methodologies. While some studies choose to disregard electronic annotations due to the common perception that they are of low quality, they nevertheless constitute over 98% of all annotations (du Plessis *et al.*, 2011) and hence greatly increase coverage of the protein universe when considered. In fact, as a means to increase coverage, in this work all annotations from the GOA project will be considered irregardless of their evidence codes.

2.5 Semantic similarity

According to Gruber (1993) the specification of a conceptualization that describes concepts and relationships used within a community is defined as an ontology. Consequently, semantic similarity can be defined as the quantity that reflects the closeness in meaning of two concepts in an ontology. However, the semantic similarity between two proteins annotated with GO terms is commonly called "functional similarity" since their functional annotation terms are being measured for their similarity.

Semantically similarity metrics for GO terms commonly employ either edge-based approaches or node-based approaches (or even sometimes hybrid approaches). As implied by the names, *edge-based* and *node-based* approaches use respectively edges and nodes (and their respective properties) as data sources. Typically, in edge-based approaches the most simple method relies on counting the number of edges between two terms on the ontology graph, thus conveying a measure of distance that can be easily converted into a similarity measure (Rada *et al.*, 1989). Hence, the shorter the distance between two terms, the more similar they

are. Using Figure 2.7 as an example the distance between *lipid biosynthetic process* and *lipid catabolic process* is 2. Alternatively, the common path technique can be employed, which is given by the distance between the root node and the lowest common ancestor (LCA) the two terms share (Wu & Palmer, 1994). In this case, the longer the distance between the root and the common ancestor, the more similar are the terms. Resorting again to Fig. 2.7, under this technique *lipid biosynthetic process* and *lipid catabolic process* share higher similarity than *biosynthetic process* and *metabolic process* since the former have *lipid metabolic process* as a LCA, which is at a distance of 2 from the root, whereas the latter have *biological process* as LCA, which since it is the actual graph root, naturally it is at a distance of 0 of itself. Node-based measures are more suitable to handle GO terms since there is not an uniform distribution of nodes and edges and also different edges convey different semantic distances. A commonly used node property is the information content (IC), which gives a measure of how specific a term is within a given corpus (Resnik, 1995). Hence, a GO annotated corpus like the one provided by GOA is well suited to take advantage of this property. The IC of a term can then be given by:

$$IC(t) = -log_2 f(t)$$

where f(t) is the frequency of annotation of term t. Consequently, terms annotating many proteins have a low IC, while specific terms only annotating a few proteins have an high IC. Considering that two terms are as similar as the information they share, semantic similarity measures making use of the IC property commonly compare it at the common ancestors that two terms have. Semantic similarity measures using IC, commonly compare it at the common ancestors that two terms have, considering that two terms are as similar as the information they share. The two most general IC-based approaches use either the most common ancestor (MICA) technique, where only the common ancestor with the highest IC is considered (Resnik, 1995), or the disjoint common ancestor (DCA) technique, which considers all common ancestors that do not incorporate any other ancestor (Couto & Silva, 2011). Popular node-based measures include Resnik's, which only considers the IC of the ancestor (Resnik, 1995), and Lin and Jiang,

and Conrath's, which consider the IC of both the ancestor and the terms being compared (Jiang & Conrath, 1997; Lin, 1998).

Protein semantic similarity is given by comparing the sets of GO terms (within each GO aspect) annotating each protein. There are two main approaches that can be used for measuring protein semantic similarity, pairwise and groupwise (Pesquita *et al.*, 2009). Pairwise approaches are based on combining the semantic similarities between the terms that annotate each protein. These use only direct annotations and apply term semantic similarity measures to all possible pairs made between each set of terms. Variations of the pairwise approaches include considering every pairwise combination (all-pairs technique) or only the best-matching pair for each term (best-pairs technique). Commonly, the pairwise similarity scores are combined by average, sum, or selecting the maximum to obtain a global functional similarity score between proteins.

Consider the example on Figure 2.8 where two hypothetical proteins, A and B, along with their GO annotations (direct and inherited) from the molecular function aspect are represented. In this example, the *all-pairs* technique would calculate the similarity for all four pairs of directly annotated terms, whereas the *best-pairs* technique would only consider the pairs *transcription factor activity – transcription cofactor activity* and *transcription factor binding – DNA binding*. The final value would then be given by the maximum, average, or sum of these similarities.

Groupwise approaches on the other hand calculate similarity directly, without applying term similarity metrics. These approaches can be grouped into three categories: set, vector, or graph. Set-based measures consider only direct annotations and use set similarity metrics, such as simple overlap. Vector-based measures consider all annotations and represent proteins as vectors of GO terms and apply vector similarity measures, such as cosine vector similarity. Graph-based measures represent proteins as the subgraphs of GO corresponding to all their annotations (direct and inherited). In this latter case, functional similarity can be calculated either by using graph matching techniques or, because these are computationally intensive computations, by considering the subgraphs as sets of

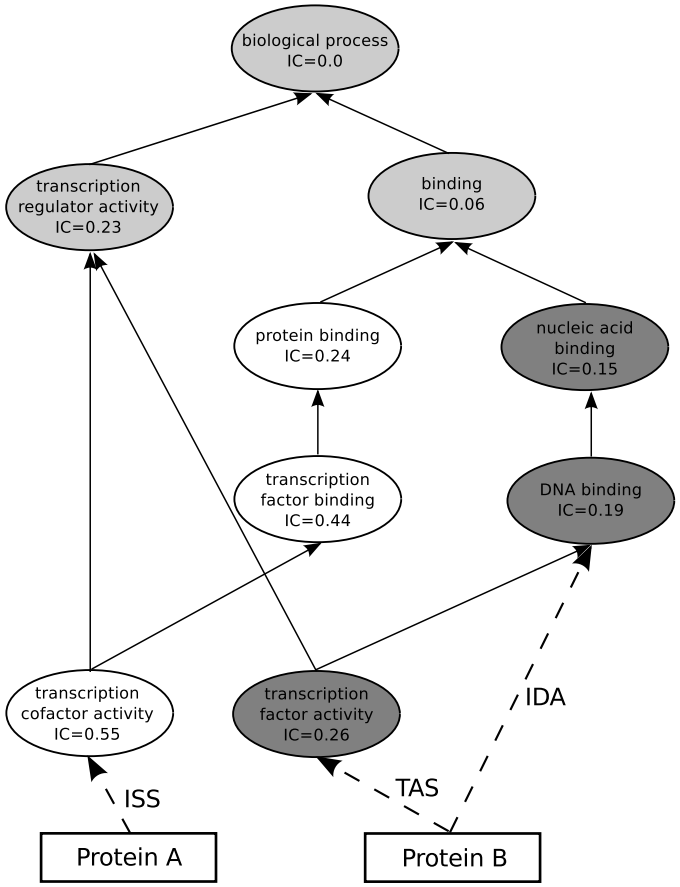


Figure 2.8: *Illustration pf graph-based semantic similarity. Full lines are GO edges and dashed lines represent annotation identified with their evidence codes.*

terms and then applying set similarity techniques. A popular set similarity technique used for this is the Jaccard similarity, whereby the similarity between two sets is given by the number of elements they share divided by the number of elements they have in total. The Jaccard similarity can be applied directly to the number of terms (simUI) (Gentleman, 2005) or be weighted by the IC of the terms (simGIC) (Pesquita *et al.*, 2008) to give more preponderance to more specific terms. For two proteins A and B, their sets of GO term annotations being GO(A) and GO(B), respectively, simUI is given by the number of terms in the intersection of GO(A) and GO(B) divided by number of terms in their union (Gentleman, 2005) as shown in Equation 2.1.

$$simUI(A,B) = \frac{COUNT\ (t_{\in GO(A) \cap GO(B)})}{COUNT\ (t_{\in GO(A) \cup GO(B)})} \tag{2.1}$$

whereas simGIC instead of using term counts, uses the sum of each term's IC in the intersection of GO(A) with GO(B) divided by the sum of each term's IC in their union as demonstrated in Equation 2.2.

$$simGIC(A,B) = \frac{\sum t_{\in GO(A) \cap GO(B)} IC(t)}{\sum t_{\in GO(A) \cup GO(B)} IC(t)} \tag{2.2}$$

Figure 2.8 illustrates this type of measure, since each node color identifies it as a term that strictly belongs to a single protein's annotations (white or dark gray) or to both (light gray). Using simUI, the similarity between the proteins would be 0.33, whereas using simGIC it would be 0.14.

The semantic similarity measures for GO terms have been developed and employed on various assessment studies (Guo *et al.*, 2006; Pesquita *et al.*, 2008). However, while a given measure can be suitable for one task it may not perform well on another. Thus, there is not one clear best measure for universally comparing terms or proteins.

2.6 Term enrichment analysis

Among the analysis operations involving GO terms, term enrichment analysis is one the most commonly used. Micro-array experiments often output lists which can represent hundreds or thousands of genes found to be differentially regulated for a given condition under study. The purpose of term enrichment analysis is then to abstract from the individual genes and focus instead on a representative set of activity terms that summarize the particular biological activity differential, characteristic of the condition being studied.

The term-for-term approach is one of the commonly used term enrichment analysis techniques. Considering the application of this technique to protein sets, for each annotation term in a study set, the purpose is to test the null hypothesis that states that there is no association between the number of annotated proteins in a set and the number of annotations of that given term, against an alternative

hypothesis of an association existing between them. That is, each set is considered to be, by the null hypothesis, just a random sample of the population. The choice of the right population (statistical background) is then paramount for retrieving good enrichment results. Under the term enrichment analysis, the occurrence of enrichment (or depletion) can be asserted resorting to commonly used statistical tests for this effect, such as the Fisher exact test, the Chi-squared test, the Hypergeometric distribution and Binomial distribution.

Huang *et al.* (2009) collected and reviewed 68 bioinformatic enrichment tools categorizing them into three different classes, singular enrichment analysis (SEA), gene set enrichment analysis (GSEA) and modular enrichment analysis (MEA). Common to these three categories is the computation of p-values which for SEA is done for each term in a list of pre-selected genes deemed of interest, whereas GSEA needs no pre-selection and has experimental values integrated directly into p-value calculation. On the other hand MEA is similar to SEA but additionally factors term-term and gene-gene relations into the p-value calculations. However, and despite the number of available enrichment tools there are still several unaddressed issues, even if we disregard issues stemming from experimental design and execution. These originate from variations in the sizes of the lists of genes, dependencies among genes or terms, annotation incompleteness and overall heterogeneity regarding specificity of annotation. And while the MEA methods try to address and even take advantage of the possible dependencies between genes or terms, issues pertaining to heterogeneous term availability or annotation distribution can still cause several problems and are still not optimally addressed.

Considering once again the ubiquitous term-for-term approach, it should be noted that the graph nature of GO leads to a statistical dependency issue. That is, for a given term annotating a certain number of proteins, at least that same number of proteins or more will also be annotated by the parental terms. Among the several strategies used to mitigate this issue, here, the Topology-based Elimination (Elim) strategy (Alexa *et al.*, 2006; Robinson & Bauer, 2011) is highlighted. This strategy consists in targeting significant leaves in an annotation graph and iteratively subtracting the proteins annotated there from parent terms up until the root term. After all terms are processed new p-values are computed for each term. Additionally, it should be noted that the computed p-values for the GO

terms under this strategy are conditioned on their children terms, and thus not independent. Therefore, direct application of the multiple testing theory is not possible. It is then preferable to interpret the returned p-values as corrected or not affected by multiple testing.

Summary

In this chapter an overview of the molecular biology of proteins was presented in order to introduce the concepts needed to understand the underlying decisions in the explored and developed methods. In addition, it was described how protein information repositories are managed from large databases dedicated to storing protein sequence data, to smaller, more specialized databases dedicated to gather protein related information focused on particular fields of interest. Some protein functional classification schemes were introduced with focus given to GO, currently the most prevalent annotation ontology for proteins. The concepts of GO-based semantic similarity and term enrichment analysis were also approached to highlight some of the advantages of using a controlled vocabulary like GO to annotate proteins.

The following chapter presents an exploration of the common functional annotation methodologies. These are classified into different types and examples of annotation systems illustrating those types are discussed and compared by segregating the annotation processes into two common stages. In addition, the chapter presents existing annotation coherence metrics and annotation management systems.

Chapter 3

State of the Art

The focus of this thesis is the assessment of functional coherence in protein sets (families). However, one of the obstacles of accurately measuring functional similarity between proteins (and coherence for protein sets) is the heterogeneity of the specificity in protein annotations. Therefore, the state-of-the-art in protein functional annotation is explored in depth and described in Section 3.1 in order to provide a better understand the most common annotation processes. Subsequently, the state-of-the-art for metrics and approaches used in the assessment of annotation coherence and annotation management, such as post-annotation improvement and error finding are described in Section 3.2.

3.1 Functional annotation

One of the most challenging tasks in genome sequencing projects is the functional annotation. This task is essential in providing biological contexts to genome sequences and thereby facilitate knowledge exchange within the scientific community. Genome annotation typically handles annotation of the most well-studied biological entities such as gene-products(protein-genes; RNA-genes) and repetitive DNA.

Annotation efforts can be analysed from several perspectives. Socially, these genomic features are commonly annotated using one of four established annotation paradigms, the "museum", "cottage industry", "jamboree" or "factory" models of genome annotation (Elsik *et al.*, 2006; Stein, 2001). The first three mentioned

paradigms are used in manual or semi-automatic genome annotation efforts. The annotation under the "museum" model relies on a small group of specialized curators. Those curators systematically annotate the genome finding and correcting errors produced by gene-predicting and functional annotation algorithms. The "museum" model due to the cost of supporting a group of specialized curators is more frequently used for model organisms that have sufficient funding, for example, the Rat Genome Sequencing Consortia (Gibbs *et al.*, 2004). The "cottage industry" model is a variation of the museum model where the specialized curators work part-time and are recruited from the ranks of post-doctoral fellows, graduate students and faculty (Stein, 2001). In the "jamboree" model a group of leading biologists from the community and bioinformaticians come together to annotate a set of predetermined entries in a given amount of time (Mazumder *et al.*, 2010). Annotation jamborees have been used for a number of early reference genome projects (e.g. Drosophila melanogaster, the mouse full-length cDNA project, the Escherichia coli K-12 and rice genomes (Kawai *et al.*, 2001; Ohyanagi *et al.*, 2006; Pennisi, 2000; Riley *et al.*, 2006)). Additionally, several community annotation databases and distributed annotation infrastructures (Griffith *et al.*, 2008; Menda *et al.*, 2008; Thornton, 2009) allow continuous input in order to produce full genome annotations. These disperse community annotation approaches can bring advantages such as focus on annotation of gene families of interest. Furthermore, the expertise of a potentially greater number of biologists can be exploited for functional annotation, additional laboratory experiments can be done for validation and a larger number of automated gene predictions can be verified more quickly. However, these annotation models involving extensive participation of specialized curators still have a relatively slow annotation output speed. Thus, the "factory" model which employs highly automated annotation methods is typically used in many full genome annotation projects. Also, in this manuscript, there is a greater focus on systems using this annotation model.

Biological entities can be annotated from different perspectives encompassing multiple features ranging from physical properties and biochemical function to interactions occurring during biological processes. It is then possible to adapt functional annotation classification into four distinct dimensions, one-dimensional

1-D	2-D	3-D	4-D
genome annotation	biological network reconstruction	compartmentalization	evolutionary scale

Table 3.1: *Summary of aspects focused within each dimension of functional annotation.*

or genome annotation, two-dimensional, three-dimensional and four-dimensional as proposed by Reed *et al.* (2006) and summarily represented in Table 3.1. A one-dimensional annotation is the commonly regarded genome annotation, consisting in the identification of genes and the assignment of known or predicted function of the gene products. In turn, a two-dimensional genome annotation will specify cellular components and provide information on how those components interact encompassing, for example, metabolite transformations, protein-protein interactions and regulatory interactions. The knowledge of these relations allows for a two-dimensional representation of information leading to the reconstruction of the biological networks. A three-dimensional accounts for intracellular arrangements or compartmentalization of cellular components, which can play an important role in their function. The fourth-dimensional annotation is a temporal dimension on an evolutionary scale, and it tracks the changes in the genome sequences occurring throughout evolution. Given the focus of this work, only protein (enzyme) functional annotations at the one-dimensional annotation level and some aspects of a two-dimensional annotation will be considered and discussed in this document.

With the increased availability of biological sequences, several systems for automatic annotation of biological sequences have been developed. In the timeframe of just one year Rost *et al.* (2003) found over 1000 results for publications pertaining to protein function prediction. However, in this document only the most representative examples will be analysed in order to illustrate the most common approaches used in automated annotation systems.

3.1.1 Annotation systems

Most automatic protein annotation systems do not actually produce *de novo* functional annotation terms. Instead, those systems commonly rely on methods for transferring annotation terms from previously annotated protein sources to other unannotated (or incompletely annotated) proteins. So, the general streamline of an automatic annotation system at a first stage involves the *identification of functional peers* on annotated corpora. Subsequently an actual *annotation transfer* stage occurs where typically terms can be either directly transferred to the unannotated proteins or be further processed to achieve more precise functional assignments. In this manuscript these two stages, *identification of functional peers* and *annotation transfer* will be considered separately for convenience of analysis.

Despite that, in proteins, structure is generally more conserved than sequence and the wider availability of sequence data over structural data allows for a potential greater annotation coverage with the former. Still, proteins having similar sequences can typically hold evolutionary proximity, and to some extent function conservation thus providing sound approximations. Hence, and given the provided coverage advantage, this work will focus on and discuss essentially sequence-based approaches to functional annotation. These approaches can be broadly grouped, as depicted in Figure 3.1, into three specific methodology types: homology-based, motif-based and genomic context strategies.

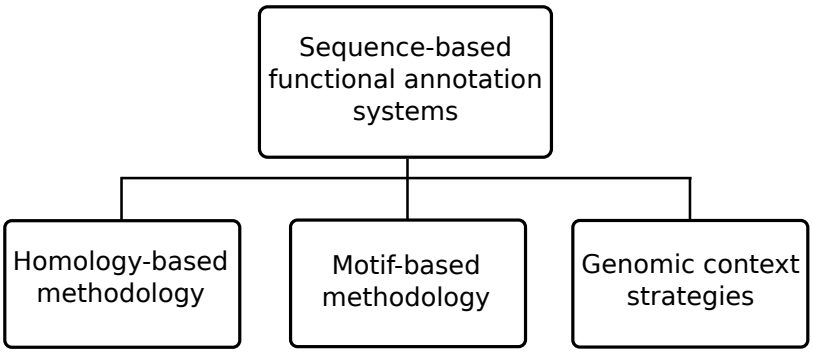


Figure 3.1: *Sequence-based functional annotation systems.*

The homology-based methodology is perhaps the most prevalent among the functional annotation systems. These methodologies generally make use of sequence alignment algorithms, such as the ubiquitous BLAST (Altschul *et al.*, 1997), to compare unannotated query proteins against annotated sequences in a database. The underlying assumption is that similar sequences are most likely to have evolved from a common ancestor and thus retained similar functions. However, high sequence similarity does not always mean function similarity (Rost, 2002) so annotation systems also employ additional techniques to handle known issues, that will be discussed further ahead.

Instead of querying unannotated sequences against databases of annotated sequences, they can instead be queried against known recurring patterns of motifs known to be associated with particular functions. Previously, in Chapter 2 some of the most commonly used databases hosting this type of information were introduced. An annotation system can then use the patterns, rules and profiles of PROSITE, the fingerprints in PRINTS, the family profiles from ProDom, the HMMs from Pfam databases or any other sequence motif type in order to perform functional inference. This is the so-called sequence motif-based methodology and can be either used on its own or often as complementary methodology in tandem with homology-based methods.

Unlike homology-based and motif-based, some other alternative sequence-based methodologies exist that do not rely on sequence comparison techniques. These can collectively be called the genomic context strategies and typically comprehend the use of gene neighbourhood (Bowers *et al.*, 2004), gene clustering (Overbeek *et al.*, 1999), Rosetta stone methods (Zhang *et al.*, 2006) and phylogenetic profiles (Gonzalez & Zimmer, 2008) either conjointly or individually. Additionally other methods such as gene expression (Zhang *et al.*, 2004) and protein-protein interactions methods (Chua *et al.*, 2006) are also used at the functional peers identification stage of functional annotation.

3.1.2 Identification of functional peers

GENEQUIZ (Andrade *et al.*, 1999; Scharf *et al.*, 1994) belongs to the first generation of fully automatic systems to provide annotations for biological sequences.

This is one of the precursor systems that rely on querying several databases for the retrieval of homologue sequences. The source databases for this system are SWISS-PROT with its supplement TrEMBL (Boeckmann *et al.*, 2003), PIR (Wu *et al.*, 2002), GenPept, EMBL and GenBank. The method employed for homologue retrieval in this system is quite simple. A number of different databases are searched with sequence alignment algorithms (using programs from the BLAST or FASTA packages) and results are ordered by decreasing similarity to sequence. However, all these mentioned databases have different free-text functional annotation descriptors for their sequences. Since the annotation of the new sequences will be dependent on the previously annotated sequences using different descriptors has the disadvantage of leading to annotation inconsistencies. Approaches that deal directly with this issue will be addressed and discussed later.

The Blast2GO (Götz *et al.*, 2008), GOPET (Vinayagam *et al.*, 2006) and GOtcha are three annotation systems, that like GENEQUIZ, follow a simple homology-based methodology. In these three systems homologue proteins (functional peers) are found using the ubiquitous BLAST sequence alignment algorithm. Abascal & Valencia (2003) in their annotation systems also rely on BLAST for their homologue identification step, however they follow that technique with a clustering algorithm in order to identify more closely related sequence groups. Given that sequence similarity not always precludes homology, some systems also use additional techniques or methodologies to ascertain proper identification of functional peers. Fleischmann *et al.* (1999) developed the system for the automatic annotation of TrEMBL which uses PROSITE signatures, thus a motif-based methodology, in addition to a previous homology-based methodology. Kretschmann *et al.* (2001) system also makes use of PROSITE signatures and additionally takes taxonomy of the organisms in consideration. The Bioverse (McDermott & Samudrala, 2003) system goes farther and besides using PROSITE it also uses BLOCKS and PRINTS for sequence pattern matching and further integrates domain and family information from several databases [Pfam, ProDom, SMART, TIGRFAMs]. Both the MaGe annotation system (Vallenet *et al.*, 2006) and the system developed by Zheng *et al.* (2005) use the homology-based approaches for functional peer identification. Additionally, the actual identification

of homologues in these two systems is done with the help of special sequence features, syntenic anchors (Mural *et al.*, 2002). These syntenic anchors are conserved non-repetitive locations between two different genomes that may correspond to stretches of contiguous genes that are likely orthologs (Vidal *et al.*, 2003).

Numerous automatic annotation systems such as Genotator (Harris, 1997), RiceGAAS (Sakata *et al.*, 2002), cDNA2Genome (del Val *et al.*, 2003), CAAT-Box (Frangeul *et al.*, 2004), BASys (Van Domselaar *et al.*, 2005), MaGe and others were designed to be part of the pipeline of genome sequencing projects. Thus, their inputs are typically the sequence strings of the cDNA clones, as provided by their respective genome sequencing projects experiments. Hence, the pipeline in these systems typically integrates several applications that execute primary tasks, such as gene finding, gene splicing and exon translation which are important steps leading up to actual protein sequence annotation. An example is RiceGAAS which is a system developed to provide automated annotation for rice genome sequences (Sakata *et al.*, 2002). Its pipeline includes applications for RNA prediction, exon prediction, splice site prediction, domain prediction and repetitive sequence detection among others. However, more relevant for this work, is that near the end of the pipeline RiceGAAS also employs a typical homology-based methodology (using BLAST) complemented with a motif-based methodology (using HMMER, ProfileScan and MOTIF). Likewise, other complete genome-oriented annotation systems like Genotator (Harris, 1997), GAIA (Bailey *et al.*, 1998), GenDB (Meyer *et al.*, 2003), cDNA2Genome (del Val *et al.*, 2003), CAAT-Box (Frangeul *et al.*, 2004) and BASys (Van Domselaar *et al.*, 2005) all use homology-based methodologies function annotation once they derive protein sequences from the initial genomic sequences. In addition, just like the Bioverse system they can use other sequence-derived features in order to validate homology assumptions. Additional systems, such as the ones by Jensen *et al.* (2002, 2003), instead of relying on sequence homology altogether, they use other features such as predicted post translational modifications (PTMs), protein sorting signals and physical-chemical properties calculated from the amino acid composition. However, it should be noted that unlike the motif-based methodology, here the features are directly derived from the input sequences and not obtained

	Homology-based	Motif-based	Genomic context strategies	Other
GENEQUIZ	X			
Blast2GO	X			
GOFigure	X			
GOblet	X			
GOPET	X			
GOtcha	X			
Abascal & Valencia, 2003	X			
Fleischmann *et al.*, 1999	X	X		
Kretschmann *et al.*, 2001	X	X		
Bioverse	X	X		
Zheng *et al.*, 2005	X			X
MaGe	X	X		X
Genotator	X	X		
RiceGAAS	X	X		
cDNA2Genome	X	X		
CAAT-Box	X	X		
BASys	X	X		
Overbeek *et al.*, 1999			X	
Zhang *et al.*, 2004			X	
Deng *et al.*, 2002			X	
Letovsky & Kasif, 2003			X	
Prolinks			X	
Chua *et al.*, 2006			X	
Jensen et al., 2002		X		X
Jensen et al., 2003		X		X
Renner *et al.*, 2000				X
Raychaudhuri *et al.*, 2002	X			X
Couto *et al.*, 2003				X
Aerts *et al.*, 2008				X

Table 3.2: *List of annotation systems and their respective functional peer identification methodologies.*

through pattern matching on external databases.

Some alternative annotation strategies can be categorized under the denomination of genomic context approaches. These approaches subsume the gene neighbourhood, gene clustering, Rosetta stone and phylogenetic profiles methods. These methods operate by identifying pairs of non-homologous proteins that co-evolve. The evolutionary pressure originates pairs of proteins that functionally collaborate and that: i) are coded nearby in multiple genomes, the gene neighbourhood method; ii) are components of an operon in prokaryotes, the gene cluster method; iii) can be fused into a single protein in some organisms, the Rosetta stone method; iv) are regularly both present or both absent within genomes, the phylogenetic profiles method (Bowers *et al.*, 2004). Protein-protein interactions and the gene expression data from microarray experiments have also been used as part of the functional peers identification methodology in some annotation systems. All of these methodologies are present in systems that aim at a two-dimensional genome annotation, that is, their main goal is typically network reconstructions. However, these methods will still be discussed, if only briefly, because one-dimensional annotations can be derived from biological process assignments or constructed protein networks.

These genomic context methods can be used on annotation systems either individually or conjointly. Overbeek *et al.* (1999) apply the gene clustering method on their system to infer functional coupling in prokaryotic genomes. Zheng *et al.* (2002) also uses a clustering method but applied on phylogenetic profiles. Using microarray mouse expression data for nearly 40,000 known and predicted mRNAs in 55 mouse tissues Zhang *et al.* (2004) were able to show that quantitative transcriptional co-expression is a powerful predictor of gene function. On the other hand, the Prolinks (Bowers *et al.*, 2004) database uses the four genomic methods described above combined to infer functional linkage between proteins through the identification of pairs of non-homologous proteins that co-evolve. Phydbac, a gene function predictor specialized in bacterial genomes, uses just three of genomic context strategies. Protein associations are generated by phylogenetic profiles, proteins evolving in a correlated manner tend to intervene in common metabolic pathways or constitute multi-molecular complexes; co-localization is

used under the assumption that genes separated by small intergenic distances are likely to belong to a shared operon for all prokaryotic organisms; Rosetta Stone methods are used to detect gene fusion events, since distinct genes of one organism that are found fused on another organism tend to physically interact (Enault *et al.*, 2005). Both Deng *et al.* (2002) and Letovsky & Kasif (2003) employ the theory of Markov random fields to infer protein functions using protein-protein interaction data and the functional annotations of protein interaction partners. Chua *et al.* (2006) also develop a method for predicting protein function based on protein-protein interacting data, the difference being that in this case transitive relations are also considered for the predictions.

Other annotation approaches to functional annotation exist that do not rely (fully or partially) on sequence. Among these, systems that employ text-mining strategies for functional annotation are also common (Aerts *et al.*, 2008; Couto *et al.*, 2006a; Raychaudhuri *et al.*, 2002; Renner *et al.*, 2000).

3.1.3 Annotation transfer

Prediction and assignment of protein function is seldom done in a deterministic way. While some general functions can be assigned deterministically to sequences, as protein function specificity rises the uncertainty of an exact assignment does also. Thus, after the identification of functional peers it is common for annotation systems to employ an additional stage where selection of the actual terms to be transferred occurs. A confidence measure is usually associated with these term transfers, which often derives directly from probabilistic features of statistical or machine learning methods employed for term selection, or arbitrary empirical confidence measures from rule-based term selection methods. In this section annotation systems will be discussed regarding the methodology used for term selection and transfer. These methodologies can be roughly grouped into three types: rule-based transfer, statistical transfer and machine learning transfer. Table 3.3 displays the categorization of annotation systems in terms of the annotation transfer methodologies they employ. Furthermore the type of annotation being transferred by each system will also be discussed.

	rule-based transfer	statistical-based transfer	machine learning-based transfer
GENEQUIZ	X		
Blast2GO	X		
GOFigure	X	X	
GOblet	none	none	none
GOPET			X
GOtcha		X	
ProFAL	X		
Fleischmann *et al.*, 1999	X		
Renner *et al.*, 2000			X
Kretschmann *et al.*, 2001	X		X
Raychaudhuri *et al.*, 2002		X	
Jensen *et al.*, 2003			X
Abascal & Valencia, 2003	X		
McDermott *et al.*, 2003			X
Yao *et al.*, 2006			X

Table 3.3: *List of annotation systems and their respective annotation transfer methodologies.*

As stated in the previous section GENEQUIZ is one among the first generation automatic sequence annotation systems. During the annotation selection and transfer stage of this system keywords are extracted and scored accordingly to their frequency in the classes generated in the process previously described. These keywords are then assigned to new unannotated sequences in each class only if their frequency is above a defined threshold (85%). This system denotes reliability through confidence classes. The underlying measures for this classification come from the normalized scores obtained with the sequence alignment algorithms (BLAST and FASTA) and are adjusted by arbitrarily chosen reliability values attributed to certain databases in order to favourably bias (SWISS-PROT > PIR > TREMBL, GenPept > EMBL, GenBank) them over the others (Andrade *et al.*, 1999). The automatic annotation developed for TrEMBL by Fleischmann *et al.*

(1999) instead of using several annotated data sources only uses SWISS-PROT, the manually curated counter-part of TrEMBL, thus implicitly using an higher quality underlying knowledgebase. Furthermore, it uses a rule-based methodology to annotation transfer whereas the annotation terms shared by all entries in a similarity group is only assigned if all the PROSITE signatures present in a group of similar sequences is also present in the target unannotated sequences.

One of the main issues with the two annotation systems just discussed is that they rely on reference databases manually annotated with free-text. This can lead to ambiguity and inconsistencies in the annotations.

Early on, Renner *et al.* (2000) directly addressed the problem of conflicting annotations by devising a system where the free-form annotations are clustered. That system was devised to handle cDNA as input and follows the already mentioned model where several different applications perform several tasks predicting sequence features in the pipeline of a sequencing project. However, this system is supported by an additional source of information, the biological literature. The approach used to reduce term inconsistency was the construction of clusters of co-occurring terms in paper abstracts. These clusters aim at capturing synonyms and related biological concepts under individual entities. As a way to avoid human bias Kretschmann *et al.* (2001) also uses a machine learning approach for the automatic generation of rules for protein annotation. In their system the C4.5 data mining algorithm is applied to the SWISS-PROT database. The decision trees constructed by this algorithm in this case are based on the associated PROSITE signatures and taxonomy of the protein sequences. As described before, Abascal & Valencia (2003) propose a system for automatic annotation based on the identification of families. The actual annotation selection algorithm targets those families for processing free-text annotations and keyword-like annotations associated to their members. The system uses co-occurrence of those keywords like the method by Renner *et al.* (2000), but produces a graph structure instead.

These systems described above, at their genesis, had to deal with the issues caused by the then predominant free-text annotation model. However, this issue was only natively addressed, in a widespread fashion, after the emergence of GO which since then has been providing a growing, unified and controlled vocabulary

of terms able to describe the characteristics of proteins. Due to the gain in popularity and functionality of GO, Jensen *et al.* (2002) promptly adapted their annotation system into a new system (Jensen *et al.*, 2003) using GO terms for functional annotations instead of the former free-text annotations. GoFigure (Khan *et al.*, 2003), GOblet (Hennig *et al.*, 2003), GOtcha (Martin *et al.*, 2004), GOPET (Vinayagam *et al.*, 2006) and Blast2GO (Götz *et al.*, 2008) are just some additional annotation systems that use GO as the annotation term source of biological sequences. Regarding functional peer identification all of them start either from cDNA (full genome annotation systems) or protein sequences for annotation and employ the commonly used homologue-based methodology with the ubiquitous BLAST as the algorithm for sequence alignment.

However, all these systems differ in their *annotation transfer* stage. GOblet is the simplest of those systems in that it actually does not perform actual annotation transfer, instead it displays all the GO terms associated to homologues found during the functional peer identification stage with their respective counts as a hint of which the most significant ones might be (Hennig *et al.*, 2003) leaving the term transfer decision to the user. GoFigure goes a step further in that it actually performs automated annotation transfer. This system constructs minimum covering graphs for groups of homologue sequences rooted at the common ancestor with most graph-depth. The GO terms are attributed a normalized score derived from the BLAST results, and an empirical cut-off is chosen to balance term assignment in order to assign neither too generic terms nor too specific (Khan *et al.*, 2003). GOtcha is very similar to GoFigure in that it scores terms based on the BLAST e-values producing a weighted composite subgraph of the GO, however unlike GoFigure it supplies annotations with a confidence measure instead of transferring all annotations above a pre-set cut-off empirical value. The GOPET system uses yet another type of approach, machine learning. In this system, GO terms associated to the sequences retrieved in the common first algorithmic step are used in conjunction with several elaborate attributes, including sequence similarity measures, such as e-value, bit-score, identity, coverage score and alignment length. Further attributes use GO-term frequency, GO-term relationships between homologues, the level of annotation within the GO hierarchy and homologue annotation quality which is calculated based on the evidence codes

provided by the gene association tables of the GO mapped sequence databases. These attributes are used as instances with a previously trained support vector machine (SVM) to assign GO terms to the unannotated sequences (Vinayagam *et al.*, 2006). The Blast2GO system is somewhat different in the sense that it essentially provides a framework for sequence annotation, where the actual GO term evaluation and annotation step is controlled by user-adjustable rules. It also incorporates GRID technology which is useful to reduce computation time especially in the functional peer identification while scanning for homologue sequences (Aparicio *et al.*, 2006).

The biological literature is still a major source of information, even if the information there contained is unstructured and not easily accessible. Hence, there are some annotation systems that tap into this resource, like the one developed by Raychaudhuri *et al.* (2002) or ProFal (Couto *et al.*, 2003). The former system operates using the abstracts in SWISS-PROT associated to homologue sequences found with ubiquitous BLAST program. These abstracts are subjected to maximum entropy analysis which provides a probability to each of the GO codes per abstract. The latter, ProFAL system also takes advantage of the graph structure of GO. However, instead of performing homology searches like most systems, the first step in this system consists in retrieving from several databases, paper abstracts associated to target protein sequences. GO term occurrences are then extracted from those abstracts and used to annotate proteins related to those documents, under the assumption that a protein related to a document and a GO term mentioned therein have an underlying biological relationship. Validation is done with an heuristic that checks matches between proteins from common families and common sets of biological properties (Couto *et al.*, 2006b). Aerts *et al.* (2008) also employ text-mining strategies in their system, however they aim for a two-dimensional genome annotation and focus specifically on cis-regulatory annotation.

Jensen *et al.* (2003) developed a method for predicting protein function for a subset of GO classes. The method itself relies on the input of protein sequences,

however it uses sequence derived protein features such as predicted post translational modifications, protein sorting signals, sequence length and physical and chemical properties that can be directly calculated from the amino acid composition. This method, much like GOPET, also uses a machine learning approach for annotation processing, however in this case protein features are used to feed neural networks instead of an SVM classifier. As an option to bypass optimization issues that can occur when handling SVMs, Yao & Ruzzo (2006) proposed a regression-based K nearest neighbour algorithm to perform gene function prediction using integration of heterogeneous data sources from microarray expression data and genomic sequence information. On the other hand McDermott & Samudrala (2003) resort to using neural networks for protein annotation, and relies not only on sequence derived information as input, but also combines several data sources from structural data to protein-protein interaction data.

3.2 Functional similarity

Usually, lists of proteins (or other gene products) are the resulting output of many high-throughput technologies. Therefore, identifying common functions on those sets of proteins and quantifying their functional relation is an important step in the understanding of biological systems. Over the last two decades functional annotation systems have been providing annotations for numerous proteins as well as other gene products. Additionally, the increasing popularity and consequent growth of the GO project has led to its prevalent use in annotation projects. Furthermore, the pervasiveness of GO also allowed the development of viable methodologies for the assessment of functional relatedness within sets of proteins, and other gene products, based on their respective functional annotations. Thus, the *functional coherence* here is defined as a measure of functional closeness (similarity) among all proteins of a given set. Given that functional similarity is derived from semantic similarity approaches over the annotation terms it is also relevant to define the concept of *annotation agreement* as a measure of annotation homogeneity for a given set of proteins. Methodologies for the assessment of functional coherence using annotations can often be based on the groupwise semantic similarity approaches previously discussed in this manuscript

in Chapter 2. The remainder of this section will be used to describe and discuss recently developed methodologies aiming at providing functional annotation coherence assessments in protein sets.

3.2.1 Annotation coherence metrics

GS^2 was devised by Ruths *et al.* (2009) for measuring gene set functional similarity based on GO terms. It was designed in order to be computationally efficient so that it could scale up well when comparing large gene sets. For that purpose, it uses a set-based approach for comparing genes where in a first stage annotation terms are ranked using a simple gene counting method. Those counts are followed by each gene being compared with the remaining genes regarding how it follows the distribution of functional annotations. This simple measure can only capture similarity trends within gene sets providing and cannot do exact assessment of similarity. Despite that, it was shown good performance when compared with the semantic similarity pairwise measure of Wang *et al.* (2007).

Richards *et al.* (2010) developed metrics for the assessment of functional coherence in gene sets based on the topological properties of GO-derived graphs. Their methodology relies on building GO subgraphs that subsume each gene set annotation (for each GO aspect), whereas each node is a GO term and each edge an is_a relationship between terms. Posteriorly those graphs are further enriched by adding genes, as a new type of nodes, associated to the original GO nodes, also additional new edges are created between GO terms whenever they share gene annotations. The original term-to-term edges are weighted using the IC difference between both terms while the new edges created after addition of the gene nodes to the graph are statistically weighted based on the total number of edges in the graph and the number of supporting genes for each particular edge. Hence, this approach handles both the issue at hand from an annotation enrichment perspective and annotation relationship perspective. In order to study these properties, Steiner trees are extracted from the graphs such that the sum of all edge lengths is minimized for all possible subgraphs. These properties are hence captured by resorting to three different metrics: *average seed degree*, *total*

length and *relative seed degree*. The *average seed degree* averages, for a full tree, the counts of the number of genes associated to the seed terms thus reflecting a global measure of enrichment. On the other hand the *total length* metric reflects the overall relatedness of functions by performing the sum of the length of all edges in a tree. The *relative seed degree* metric just combines the two above described aspects as a ratio. The methodology performs well, but like other GO-evaluation methodologies, it has its metrics dependent on the gene annotation status.

Diaz-Diaz & Aguilar-Ruiz (2011) approach the problem of functional coherence in gene sets by considering that each gene can encode several proteins with different functions. From there, for each gene set, in their methodology only the most common and specific function is chosen as the most globally cohesive function. In their methodology genes are represented as sets and a gene-representation similarity is calculated based on the GO-structure. They proposed a simple counting edge-based measure ratio relationship that aims at equating both gene relatedness and specificity. The final GO-based functional dissimilarity (GFD) measure is just the minimum of dissimilarity possible for all representations of a given set of genes. Like before this measure also depends on the completeness of the annotations used in order to provide accurate measurements. Furthermore, by considering only the most common and specific function in a gene set the authors are effectively discarding potential non-related functions that would cause noise, however at the cost of disregarding multi-functional associations in gene sets.

On the other hand, and despite not being exactly a system for measuring functional coherence in gene sets, RuleGO (Gruca *et al.*, 2011) provides a service that statistically compares and characterizes two disjoint gene sets. Underneath it runs a rule-based system that incrementally goes through the list of GO terms annotating the two input gene sets and verifies at each step if a new co-occurrence rule can be created. Much like the typical gene enrichment systems, this system also performs over-representation tests on the rules created and only rules under a given statistical significance (after multiple testing correction) threshold are considered. The end results are multi-attribute rules containing annotation terms and respective support statistics and evaluation parameters that can be used in the characterization of the disjoint gene sets. In this methodology rules are

evaluated by *length* (number of genes in a rule premise) representing support and *depth* (normalized sum of the GO graph levels where terms in the rule appear) representing specificity and an additional rule quality measure.

A different approach is taken by Xu *et al.* (2011) where functional coherence in gene sets is assessed with the help of the biological literature. In their approach term-by-gene matrices are constructed where the entries in the matrices are derived from weighted frequencies of the terms across a collection of abstracts (biological literature). The genes are then represented as vectors and the similarity between them is calculated by the cosine of the vector angles. Thus, a pair of genes would have a cosine score of 1.0 if they shared the exact same abstracts in the collection. Gene sets in this method were deemed functionally coherent when cosine values above a given threshold (0.6) were often found with significances measured by a statistical test (Fisher's exact test). This threshold was chosen based on the distribution of similarity cosine scores in 1,000 random gene sets. Hence, functional coherence here is given essentially through literature support thus making the method sensible to the quality of the document corpus used. The method was able to obtain results similar to the ones produced by another literature-based functional coherence assessing method (Raychaudhuri, 2003).

Since functional annotation quality is paramount Yang *et al.* (2010) developed a system to provide an annotation confidence score for genome annotations. The system operates on the basis of a genome comparison approach whereas annotations in a target genome are scored in comparison with gene annotations in a reference genome. The gene alignments across genomes are made via the BLAST tool with adjustments for expect number of genes (different organisms have different gene counts) and phylogenetic distance (closer genomes typically share more genes than distant ones). However, actual annotation similarity is derived from free-text annotations which are converted into word vectors enabling the calculation of a simple cosine similarity measure. Both sequence similarity and annotation similarity are combined into a single metric by applying statistical techniques.

3.2.2 Annotation management

The GARNET (Rho *et al.*, 2011) system while also not being aimed at providing unified functional coherence scores for gene sets is nevertheless useful in their characterization. The system allows the management of gene sets and annotation retrieval from multiple sources (as determined by the user). Typical statistical enrichment tests can be applied in order to determine the most statistically significant annotation terms, the difference here being that there can be multiple annotation sources that are handled in an integrated way. Additionally the system provides visualization tools that allow to analyse relationships between the genes.

On the other hand REVIGO (Supek *et al.*, 2011) is a system specially focused on visualization methods such as scatterplots, graph-based visualization, tree maps and tag clouds. The system calculates all pairwise semantic similarities for a set of input GO terms using the resulting matrix to perform a clustering procedure conceptually similar to the hierarchical agglomerative clustering methods. This system also relies on the GO terms being paired with p-values (or another user defined enrichment score) as part of the input, thus it can be considered as a post-processing annotation term enrichment system. The clustering procedure is applied with the intent of reducing redundancy and finding single GO term representatives for each cluster.

Despite providing no visualization methods GO Trimming (Jantzen *et al.*, 2011), much like REVIGO, also approaches the problem of redundant GO annotation terms in lists of enriched terms. It operates using a two pass methodology, where in the first pass statistically significant terms are connected to all terms sharing a common path by means of labelling with common identifiers. On the second pass, redundant terms are removed from the list, according to two approaches. In the strict approach completely redundant terms are removed, while on the *soft trimming* approach a parent term contains up to 40% genes additional genes in relation to a child term that parent term is removed. The soft trimming threshold was chosen arbitrarily based on results obtained through experimentation but it can be user adjusted.

On the other hand, GO-Chase II (Park *et al.*, 2011) aims at mitigating more than just semantic redundancy in annotation terms. The system corrects three types of semantic inconsistencies: the already mentioned semantic redundancy, derived from genes annotated with more specific terms also being annotated to more generic parent terms; biological domain inconsistencies by use of species-specific terms; and taxonomy inconsistencies where taxonomy-restricted terms are used in different taxa. For that purpose it relies on a curated database built from semantic error knowledge manually extracted from the GO-annotations from 27 of the major biological databases. In addition, it provides the typical statistical analysis of GO term enrichment tools.

Summary

In the present chapter several functional annotation systems were discussed illustrating the most common and recently proposed annotation approaches. The sequence-based annotation systems were categorized into three types: homology-based methodologies, motif-based methodologies and genomic context strategies. Regardless of their type, several example annotation systems were compared with each other by segregating their pipelines into two different stages: *identification of functional peers* and *annotation transfer*. Each of the functional coherency module and annotation extension module from the framework proposed in this thesis also overlaps considerably with these two stages. In addition, methodologies for the assertion of functional coherence and annotation management and visualization are also described in order to contextualize this thesis with the current state-of-art in those fields.

In the following chapter, an exploratory analysis of the GO annotation space of a specialized protein database is presented. Following that, the proposed completeness measures and two novel coherence metrics derived from semantic similarity and term enrichment analysis are also described and assayed along with other metrics and their results discussed. The chapter is finalized by the description of the proposed annotation extension module along with validation assay to test its viability.

Chapter 4

Functional Annotation Analysis

In order to validate the hypothesis that viable extension of functional annotations within partially annotated protein families is possible through the used of appropriated functional coherence analysis, several tests and assays were performed. Initially, an exploratory analysis of the annotation space of the proteins in a specialized protein database (CAZy) was conducted as described in Section 4.1. Following that assessment, Section 4.2 describes the testing of a range of similarity metrics, applied to functional coherence analysis. Those metrics were tested against protein families as they were increasingly altered by replacing original proteins with random ones in order to study their behaviour and resilience. Finally, Section 4.4 describes the approach and respective testing and validation of the protein annotation extension module. Additionally, the viability of the whole proposed modular framework is essayed incorporating both the functional coherence and the annotation extension module.

4.1 Protein Annotation Space

4.1.1 Annotation Exploratory Analysis

In this work, the CAZy database is used as a case-study for the assertion of coherence of the functional annotation in enzyme families. A snapshot of the CAZy database (version 7; 2010) was used for an exploratory analysis of its annotation space. This snapshot consisted of 290 protein families covering five

classes of enzymatic activities: Glycoside hydrolases (GH), glycosyltranferases (GT), polysaccaride lyases (PL), carbohydrate esterases (CE) and carbohydrate-binding modules families (CBM).

The state of annotation (and its coherence) within each of the CAZy families was initially measured by applying a pairwise semantic similarity to each pair of GO annotated proteins within each family. The semantic similarity measure used was simGIC (Pesquita *et al.*, 2008) as it had previously shown to provide a good resolution power (Pesquita *et al.*, 2009).

This assessment was limited to using just CAZy entries with UniProt identifiers because these were the only entries found to be directly linked to GO term annotations. The total CAZy family sequence space with UniProt identifiers in the analysed snapshot comprised of about 82,000 sequences, but just around 71,000 (86%) entry sequences having GO *molecular function* annotations were *de facto* used in this analysis. Only the *molecular function* term annotations were considered because the aim of this work lies closer to studying one-dimensional annotation (as proposed by Reed *et al.* (2006)) at the molecular functional level in enzymes.

Figure 4.1 shows the distribution of sizes (in number of UniProt entries) of the CAZy families with less than 1000 entries (only 15 families had more than 1000).

For a preliminary analysis, families were randomly selected from each one the five highest peaks in the histogram but assuring that all of the five different enzymatic activities classes from CAZy were represented in the selection. In order to provide a quick interpretation of the results, graphics (histograms) as the ones shown on Figure 4.2 were plotted. These histograms represent the frequency of pairs of proteins belonging to a family scoring within a certain semantic similarity range (as measured by simGIC). For the randomly selected families the coverage of the GO *molecular function* terms over the number of UniProt entries per family and over the total number of all entries per family can be seen in Table 4.1.

Following the previous described method for the random (but partially targeted) exploratory incursion, a second small set of well studied families (GH5, GH13 subfamily 2; GH14, GH15, GH70 and GH74) and also the CE and PL families (since they have a smaller number of families; see Table 4.2) were chosen

Family	UniProt entries	%GO annot. over UniProt	%GO annot. over total
CE12	53	100	31
PL9	55	60	23
CBM9	55	100	41
GT56	57	98	34
GH84	60	40	12
PL3	136	99	37
CE5	150	99	31
GH92	150	27	8
CBM3	151	100	35
GT6	161	100	25
CBM50	347	52	20
PL1	331	78	7
GH7	351	100	38
GT25	359	58	23
CE10	378	86	25
CE9	576	98	37
GT35	701	99	33
GH19	744	94	31
CE1	944	51	18
GT9	1042	98	40
GH1	1146	99	29
CBM48	1392	96	33

Table 4.1: *Randomly sampled families from each of the five highest frequency peaks from the histogram in Figure 4.1. For each family, the number of UniProt entries, the percentage of those entries annotated with GO molecular function terms, and the percentage of those annotation over the complete (UniProt plus non-UniProt entries) family size is presented.*

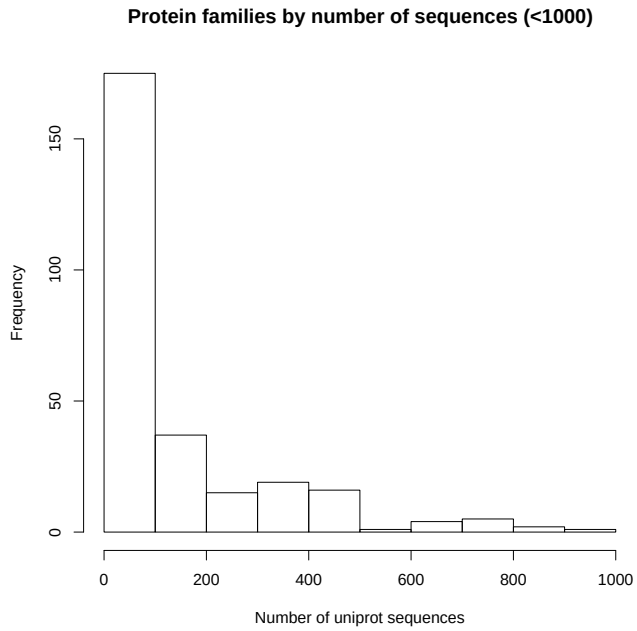


Figure 4.1: *CAZy (2010) family size distribution. Each bar shows the number of families (frequency) with a given number of sequences.*

for a more detailed analysis. Furthermore, the small set of well studied families was additionally submitted to being measured with the implemented metrics *GO score* and *GO occurrence* previously described by Bastos *et al.* (2007).

4.1.2 Results and Discussion

The generated semantic similarity (histogram) profiles of the sampled CAZy families provide a preliminary insight into their (annotation) functional coherence. As expected, all the (four) randomly sampled CBM families show varying degrees of semantic similarity (CBM3 family shown on Figure 4.2 a)). This was expected since these families comprise members that often associate themselves to other carbo-active catalytic modules in the same polypeptide and can target different substrate forms depending on different structural characteristics Cantarel *et al.* (2009). Hence, functional variety was already expected for families with this CAZy class.

According to the semantic profiles, the most coherent families were shown to

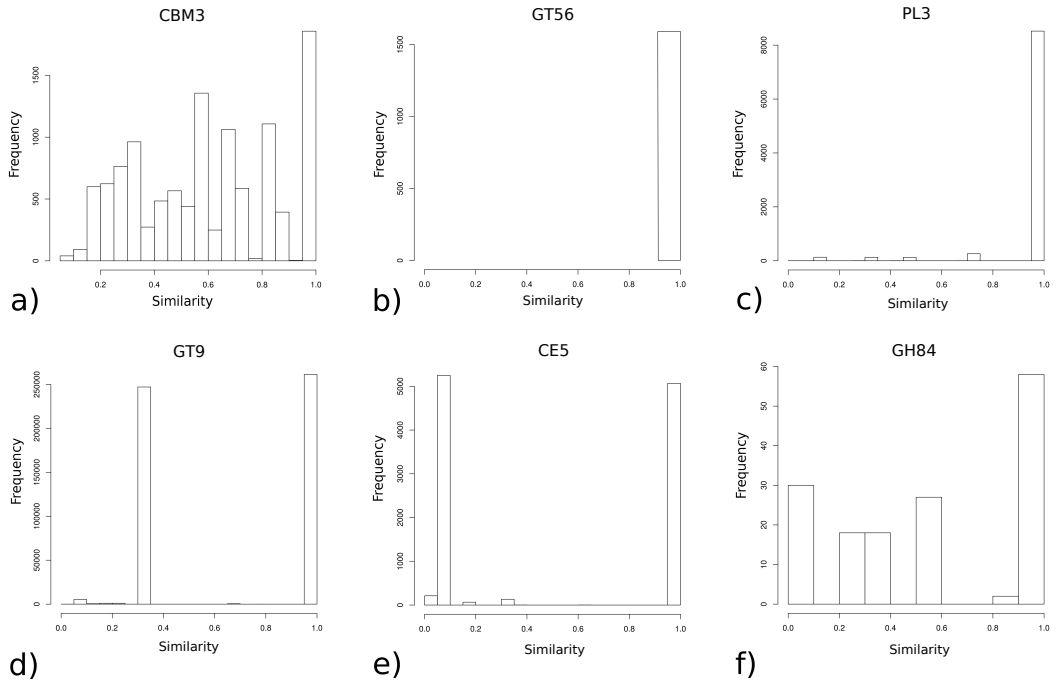


Figure 4.2: *Frequency distributions of semantic similarities between pairs of proteins in a CAZy (2010) family. Plots for family a) CBM3, b) GT56, c) PL3, d) GT9, e) CE5 and f) GH84.*

be CE12, GT56, PL3 and GH7 (GT56 and PL3 shown at Figure 4.2 b) and c) respectively), being that GT56 scored a perfect semantic similarity of 1 for all its pairs of (UniProt) proteins. This occurs because all of the considered proteins from this family are annotated to the same high informative term, *fucosyltransferase activity*. The family PL3 yielded similar results (Figure 4.2 c)) since most of its proteins are annotated with the term *pectate lyase activity*.

However, most of the sampled families (GH92, GT6, CE5, GT25, PL1, CE9, GT35, GH19, GT9 and GH1) showed a configuration where two peaks of similarity arose, one at the far right of the histogram and another one before the 0.5 semantic similarity threshold. A couple of graphics belonging to two of these families are depicted here, one on Figure 4.2 d) for family GT9 and the other on Figure 4.2 e) for family CE5. In this situation, for the case of the GT9 fam-

Enzyme class		Uniprot entries	% annotations (molecular function)
GH	(115)	39,930 (36%)	91
GT	(92)	34,947 (35%)	83
CE	(16)	5,206 (35%)	82
PL	(21)	1,052 (18%)	76
CBM	(58)	6,026 (32%)	85

Table 4.2: *The five CAZy (2010) activity classes (and number of families per class) with the corresponding number of UniProt entries per class (and respective percentage), and percentage of UniProt entries in a class that are annotated with GO molecular function terms.*

ily, most of the proteins are annotated with the term *transferase activity* while only about half of those are also annotated with a more specific term *transferase activity, transferring glycosyl groups*. Again the same behaviour is observed in family CE5, with most of the terms being annotated with the term *hydrolase activity*, and only half of them having also a more specific *cutinase activity* term annotated to them.

Family GH84 (Figure 4.2 f)) as can be seen on Table 4.1 is an example of where GO ontology still offers poor coverage. Only 40% of the Uniprot entries of this family have GO *molecular function* annotations and that covers only 12% of the total entries in the family. Although for this family there are only 24 proteins that contain GO *molecular function* annotations, the functional specificity of each annotation varies greatly. Hence, in Figure 4.2 f) we can see peaks at five different similarity levels. This, however, does not mean lack of family coherence but instead it means there is an uneven distribution of annotations regarding term specificity.

Even if GO annotations still lack in specificity, on average they cover 77% of the UniProt entries in each family. However, if we consider all the entries in each family, and not just those from UniProt, our sample data presents a coverage ranging from 7% to 41%, resulting in an average of only 28%. This means there is still a high percentage of data on the CAZy families that can be potentially explored (and undergo annotation extension).

Among the different CAZy classes, the PL class, is the class better characterized by the Glycobiology community. However, on close inspection, within the CAZy sequence space they present the lowest GO term coverage, with only 18% of UniProt entries having GO *molecular function* annotations (Table 4.2) associated to its families. Besides the small range of coverage, when analysed individually, most PL families are annotated to more conservative generic GO terms. Figure 4.3 shows the frequency distributions of semantic similarities between pairs of proteins in three (of the twenty one) PL families, PL4, PL7 and PL10. In family PL4, only the term *carbohydrate binding* stands out as the most informative common ancestor, revealing little about the molecular function of its proteins. As for families PL7 and PL10, the term *lyase activity* stands out as being the most informative common ancestor for both families. This term is certainly specific enough to classify a sequence as belonging to the PL class, however, it does not contain additional functional information to properly distinguish between different families within the PL class.

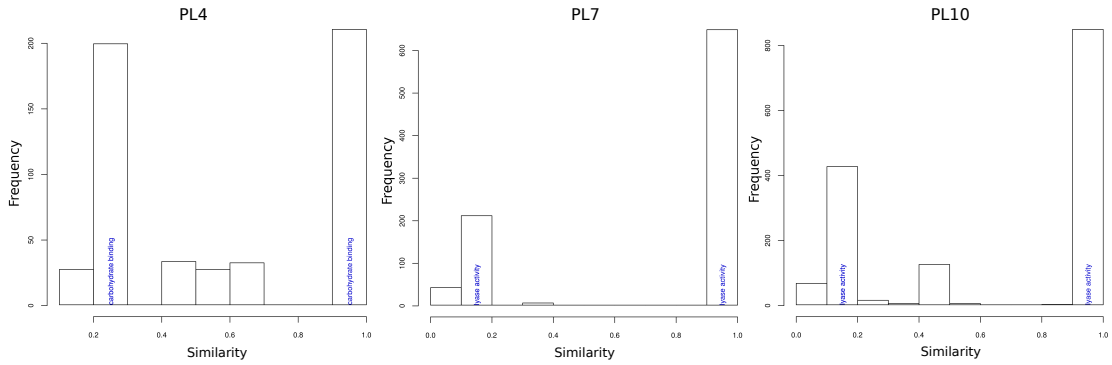


Figure 4.3: *Frequency distributions of semantic similarities between pairs of proteins in CAZy (2010) PL families. Plots for family a) PL4, b) PL7 and c) PL10*

On the other hand, Carbohydrate Esterases (CE) are known to be functionally promiscuous, in the sense that they can act on a wide range of substrates. Comparing with the PL specificity, among the total sixteen CE families only four of them were dominated by very generic GO term annotations. The rest of the CE families presented more specific *molecular function* annotations. Figure 4.4

shows three CE families, CE1, CE4 and CE11 where each one has a different most informative common ancestor dominating the functional landscape.

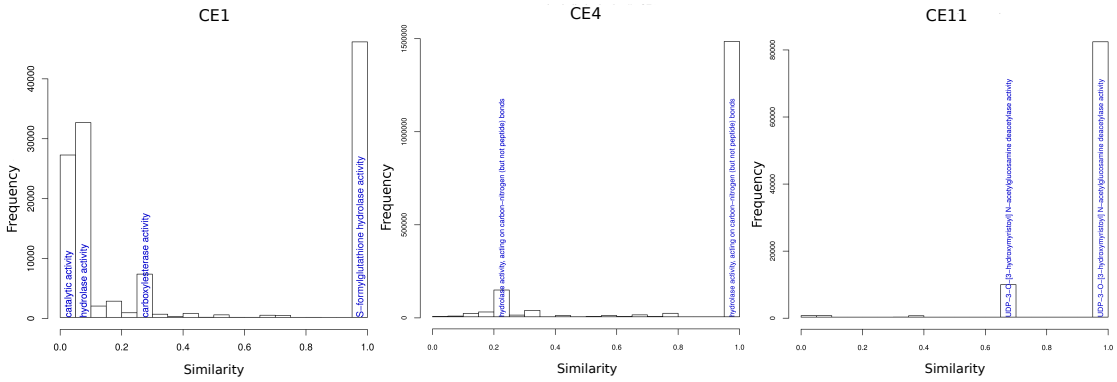


Figure 4.4: *Frequency distributions of semantic similarities between pairs of proteins in CAZy (2010) CE families. Plots for family a) CE1, b) CE4 and c) CE11*

Set of well studied CAZy families

The better studied GH5, GH14, GH15, GH70, GH74 families and GH13 subfamily 2 were examined with increased detail, and also with the additional help of the aforementioned *GO score* and *GO occurrence* metrics.

It is worth noting that some CAZy families have a computationally prohibitively large number of members to be analysed as a whole, in this manner. Furthermore, functional uniformity is frequently expected to happen only at the sub-family level. Family GH13 is a case of large family that has recently been subdivided into subfamilies (Stam *et al.*, 2006). Hence, for the current analysis only subfamily 2 of family GH13 was considered. The semantic similarity histogram for GH13 subfamily 2 is shown in Figure 4.5 a). Most noticeable is that for most pairs of proteins the semantic similarity of their annotated terms falls into ranges below a score of 0.5. On close inspection, the less informative term *cation binding* is the most prevalent, and is the most informative common ancestor for most of the ranges. The additionally used measures, GO score and GO occurrence convey an idea of specificity and agreement of GO term annotation, respectively. The small value associated to the GO occurrence (0.07) shows

GO term name	occ	score
catalytic activity	42	0.029
molecular_function	42	0.000
transferase activity	31	0.079
carbohydrate binding	33	0.256
binding	42	0.034
cation binding	42	0.129
ion binding	42	0.120
transferase activity, transferring glycosyl groups	17	0.114
cyclomaltodextrin glucanotransferase activity	17	0.303
transferase activity, transferring hexosyl groups	17	0.135
alpha-amylase activity	2	0.027
amylase activity	2	0.026
hydrolase activity, hydrolyzing O-glycosyl compounds	2	0.013
hydrolase activity, acting on glycosyl bonds	2	0.013
hydrolase activity	3	0.007
calcium ion binding	1	0.008
metal ion binding	1	0.003

Table 4.3: *GO molecular function term annotations for the CAZy (2010) sub-family GH13-2, with their respective frequency and score.*

how unevenly the annotations are distributed along the 4367 members of this sub-family. The GO score (0.15) reveals also that the most frequent and specific terms do not have an high information content (IC). As can be seen on Table 4.3 more specific terms are annotated to sub-family GH13-2, however the GO score is based on the most informative of the terms with most annotation occurrences. Hence, by itself the measure GO score is not conclusive, for instance, the size of the family and percentage of family members covered by annotation must also be considered.

Similarly to sub-family GH13-2 there are also three ranges on family GH70 (Figure 4.5 b)) annotated with the more generic *cation binding* term, however family GH70 is dominated by the *dextrasucrase activity* term. As the GO occurrence of 0.55 shows around half of the GO term annotations are distributed evenly across the family members. Also the high frequency of the *dextrasucrase activity* term gives the family a GO score of 0.32 thus raising its specificity in

relation to the former family. The results of both these families suggest that the multi-domain proteins bring noise into the functional coherence measures making the assertions unclear and more complex to determine.

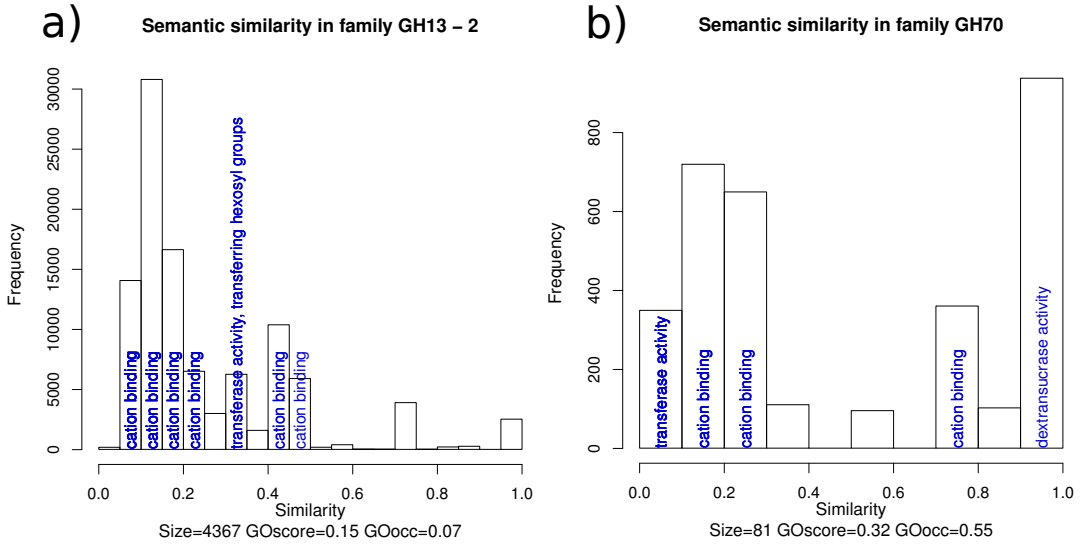


Figure 4.5: *Frequency distributions of semantic similarities between pairs of proteins in CAZy (2010) GH families. Plots for family a) GH13-2 and b) GH70.*

The families GH14, GH15 and GH77 are mostly composed by single-domain protein sequences. The frequency distributions of semantic similarities between pairs of proteins of these families can be seen on Figure 4.6. According to the graphic representation of family GH14 (Figure 4.6 a)) this is the most functionally coherent of these three families. GO occurrence supports that with a score of 0.77 and the most informative common ancestor has high IC granting this family an high GO score (0.64) confirming its specificity. Families GH15 and GH77 although having smaller GO measure scores also present profiles similar to that of GH14. The mostly mono-modular nature of these families and the specificity of its GO term annotations was confirmed by a CAZy curator.

These results show us that most of the sequence space in CAZy is still bereft of GO annotation. Additionally, even where GO annotations exist there are families where that annotation is scarce or currently only goes up to generic annotation terms. Therefore, there is still plenty of annotation growth space, both in scope and depth within the CAZy database families.

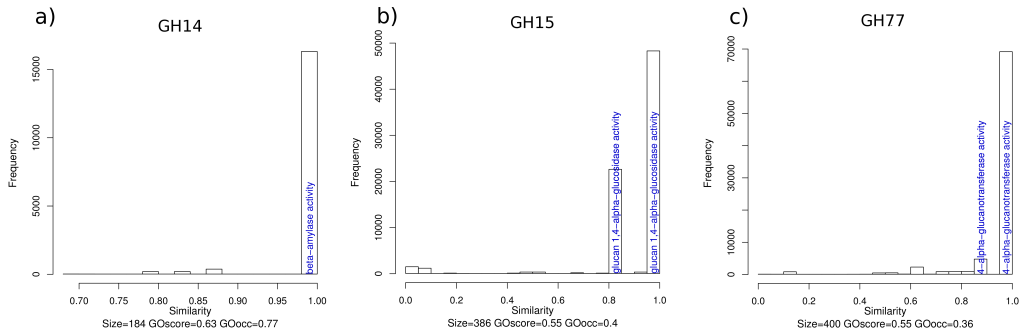


Figure 4.6: *Frequency distributions of semantic similarities between pairs of proteins in CAZy (2010) GH families. Plots for family a) GH14, b) GH15 and c) GH77.*

Furthermore, the previously presented methods only provide a small degree of insight on the annotation state of these families. Thus, new methods and tools are required both to measure annotation coherence in these families and also guide and propose annotation extension for them.

4.2 Annotation Metrics

Ideally, the functional annotations over a given protein set should allow us to infer biological relationships from such set. In order to achieve that, it is convenient to have metrics that enable us to compare how similar (or dissimilar) annotations are within a given protein set. However, considering the GO DAG structure it becomes apparent that measuring functional relatedness via annotation is not a trivial matter. Therefore, in order to make such assertions regarding functional relatedness, it is helpful to consider three annotation aspects: *completeness*, *coherence* and *agreement*, which are discussed below.

4.2.1 Completeness

Any set of functionally related proteins where not all proteins are annotated at the same specificity level can be considered to suffer from a form of *annotation incompleteness*. Figure 4.7 **a)** illustrates such a situation. For an hypothetical set of one hundred proteins, only one of the hypothetical annotation terms (besides the root) annotates all the proteins in that set. Inspecting the nodes down the

graph, as they get further away from the root term, it can be seen that the number of annotations dwindles until it reaches the leaf terms. And while any given protein does not need to have its most specific function represented by a leaf term, it is unlikely that a very generic term (like a direct child of the root term) is a full descriptor of its activity.

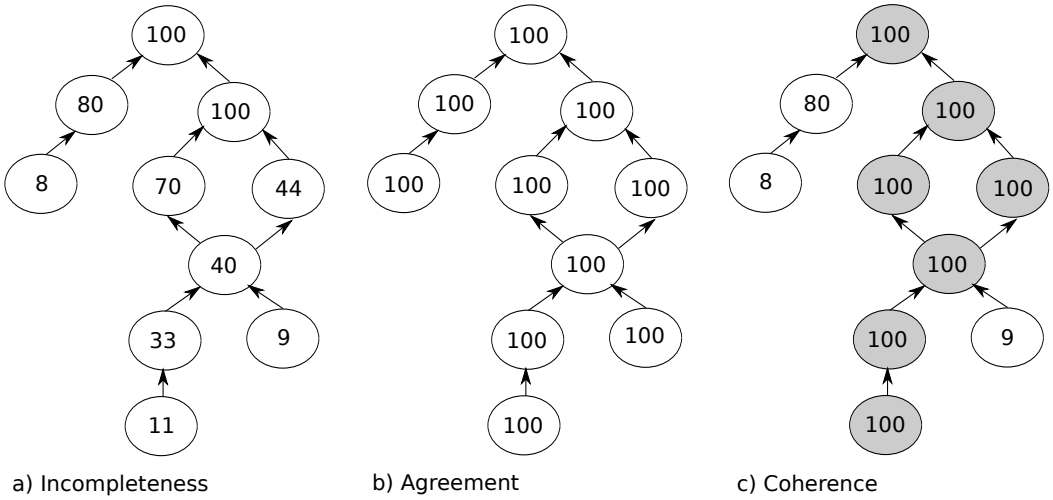


Figure 4.7: *Hypothetical GO graph where terms are represented by nodes where the number within is the number of proteins (of a given set of 100) annotated to that term. There are three situations represented: a) annotation incompleteness, b) annotation agreement and c) annotation coherence.*

This kind of annotation incompleteness can derive from the fact that different protein annotation methods are used, which provide different degrees of annotation confidence. Therefore, annotation heterogeneity is created accordingly to the annotation confidence level given by each annotation method. For instance, a large part of the automatic annotation methods typically create more generic annotations. On the other hand, manual curation is more likely to lead to more highly specific annotations. Additionally, the inherent research bias towards the more intensively studied model organisms and biological processes can also help further this state of incompleteness.

Two naive annotation completeness measures are proposed in this thesis, *Leaf-completeness* and *IC-completeness*. For any given protein Set to be considered

"fully" annotated by the *Leaf-completeness* metric, every protein in the set has to be annotated to at least one of the leaf terms from its annotation graph. The *IC-completeness* is a similar metric, whereas each protein in the set has to be annotated with a term with IC above a given user-defined threshold, instead of a leaf term.

4.2.2 Agreement

Annotation agreement can be defined as the fraction of annotations that are shared in a set of proteins. Therefore, the greater the amount of shared annotations the greater is the annotation agreement. Figure 4.7 **b)** illustrates an hypothetical full annotation agreement situation. In this situation, each one of the one hundred proteins is annotated to the same exact annotation term set and thus that hypothetical set achieves maximum or total annotation agreement. However, this is a very naïve metric that is also overly sensitive to annotation incompleteness and even small amounts of noise.

4.2.3 Coherence

Naturally, a set of proteins having a total annotation agreement is also functionally similar, to the extent of its most specific annotation terms. On the other hand, functional similarity may not need to be so strictly defined. Furthermore, due to the above mentioned incompleteness issue and the multi-functional nature of proteins, when measuring functional similarity through annotation, it may be useful to consider just some of the annotations as being functionally characteristic of a given protein set. Furthermore and for the purposes of this work, the concept of annotation coherence is further refined and defined as the fraction of shared annotations that are most relevant and characteristic in a given protein set. Figure 4.7 **c)** illustrates an hypothetical full annotation coherence situation, where the grey shaded nodes represent the functionally more relevant terms, or the central functional cohesiveness of that set. However, a single metric is too reductive in assessing these (and other) different aspects of annotation that can

dictate the functional coherence of the annotation space in protein sets. Therefore, in this thesis, a set of metrics and respective interpretation strategies are proposed for exploring protein annotation spaces.

When it comes to capturing the relationship between functional and sequence similarity, the different semantic similarity metrics often present a similar behaviour, with the main distinction among them being their resolution. In a study, comparing several GO-based semantic similarity metrics (Pesquita *et al.*, 2009), the previously presented graph-based measure simGIC, was found to be the overall best performing measure, consistently showing a high resolution (and providing about 19-44% increased resolution over the simUI metric). Both simUI and simGIC metrics were used here for assessing functional coherence and establishing similarity baselines.

As previously mentioned, in the Diaz-Diaz & Aguilar-Ruiz (2011) methodology only the most common and specific function of a set is chosen as the most globally cohesive function. In this thesis it is also assumed that not all functional annotations in any given protein set (family) should characterize that set. On the other hand, considering the frequent multi-functional nature of proteins, in this thesis, a set of annotation terms are selected in each protein set or family as being its functional characteristic core. Therefore, the strategy employed in this thesis to isolate the functional characteristic cores in protein families was to resort to term enrichment analysis. In particular, a Python implementation of the ubiquitous and previously described term-for-term enrichment approach was developed. Since in this work most of the study sets are small, and with several terms having low expected frequencies (up to five expected observations) the Fisher exact test was used to determine enrichment. The statistical evidence of enrichment was then postulated on the basis of the p-values calculated by the Fisher's exact test being smaller than the chosen statistical significance (alpha). As previously explained, the Elim method mitigates the statistical dependencies between nodes downplaying ancestor nodes, and thus was also used on this work (for an alpha = 0.01). This is a desired effect, since (for a similar level of annotation quality) a more specific annotation is preferable to a general annotation. Therefore, the Elim method favours leaf terms found to be significant and at the same time removes proteins annotated to significant children terms from the

parent terms annotation counts, which in turn attenuates the children's influence on the parental terms.

Therefore, the two novel developed functional coherence metrics, mUI and mGIC were based on a term-for-term enrichment analysis and the semantic similarity metrics simUI and simGIC as described by Algorithm 1.

Algorithm 1 Pseudo-code for calculating mUI and mGIC

```
1: INIT annotationGraph and annotationGraph'
2: FOR each term IN annotationGraph
3:    EXECUTE enrichment analysis of term
4:    IF term enriched
5:       annotationGraph' <- term
6: ENDFOR
7: mUI <- compute simUI of annotationGraph'
8: mGIC <- compute simGIC of annotationGraph'
```

The annotation graph for a protein set (family) being measured is generated (line 1). For each term (line 2) in the annotation graph enrichment analysis is performed (line 3) a term-for-term (with Elim adjustment) is performed as previously described. If a term is found to be statistically enriched (line 4) it is added to a derived annotation graph (line 5. When both annotation graphs are processed (line 6) the simUI and simGIC are applied to the shadow graph (annotationGraph') resulting in the values for the mUI and mGIC metrics, respectively (line 7 and 8). For all metrics (simUI, simGIC, mUI and mGIC), the global set results were obtained by averaging all the term pairwise results within each protein set.

4.3 Metrics assays

4.3.1 Coherence and completeness resilience assays

An experiment was designed to study the resilience of the similarity metrics. For each tested protein family (where higher similarity is expected), increasing amounts of proteins from 10% to 100% (by increments of 10%) were replaced by random proteins as depicted in Figure 4.8. Therefore, for each set (family) the

similarity was expected to degrade and the values reported by each tested metric also behave accordingly. For each of the discrete levels of noise one hundred iterations were run per family and the similarity results obtained at the end of each iteration were then averaged. At each iteration, both the original family and the noise source were randomly sampled for the replacement proteins, for the proteins to keep and the proteins to introduce, respectively.

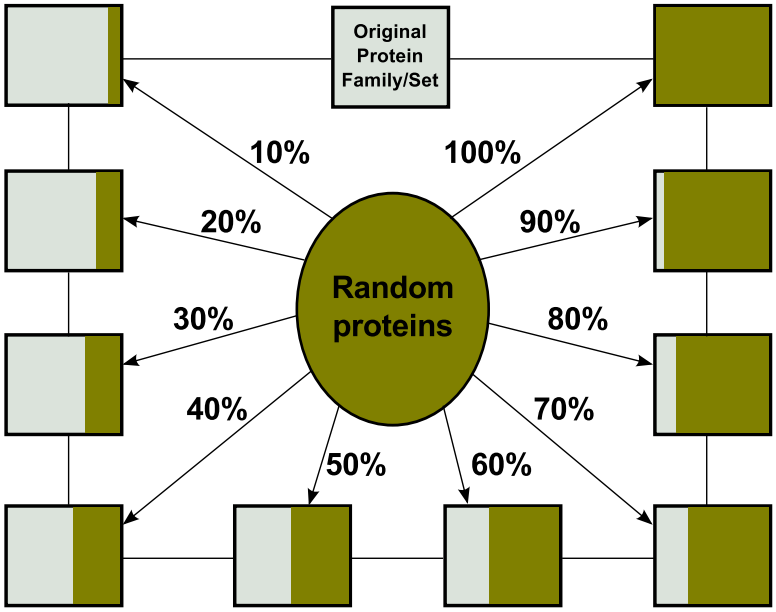


Figure 4.8: *Protein family random replacements. For each discrete percentage of noise the corresponding amount of proteins was replaced by the same random amount.*

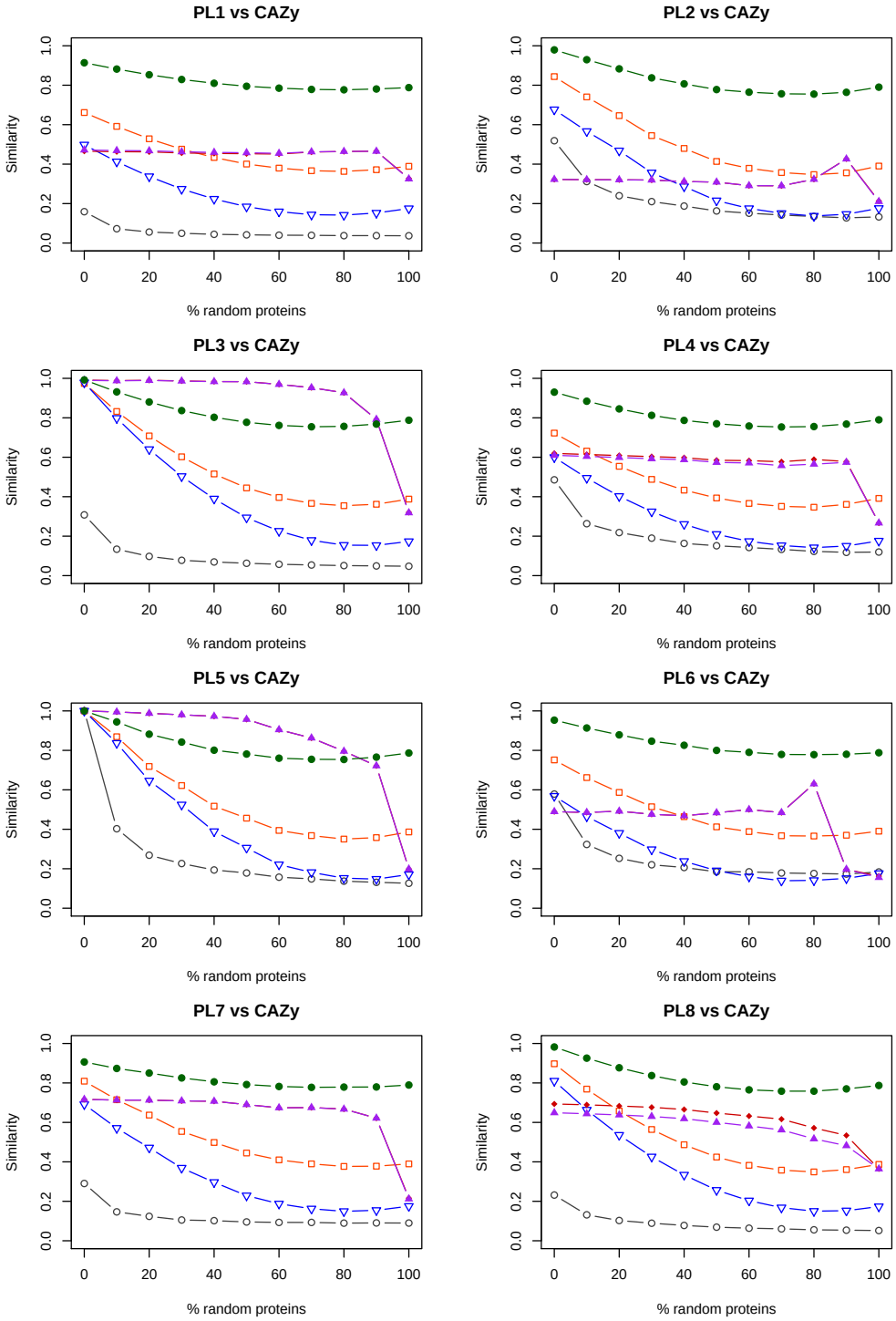
The PL families PL1 to P12, PL16, PL17 and PL22 from the CAZy database were tested against the Agreement, simUI (Gentleman, 2005), simGIC (Pesquita *et al.*, 2008), mUI, mGIG and GS^2 (Ruths *et al.*, 2009) metrics. These resulting average scores are shown in Figure 4.9 as plots of similarity (functional coherence) as a function of the percentage of introduced random proteins. For this batch of tested families the source of random proteins was of the CAZy database sequences (comprising 119,590 proteins and excluding the ones from each respective tested family). The complete tabular results are available in Appendix A. In addition to being measured with the similarity metrics, the aforementioned sets were also submitted to measurement with two completeness metrics, Leaf-completeness and

IC-completeness. The results for these two completeness metrics are made available in Appendix B in tabular form.

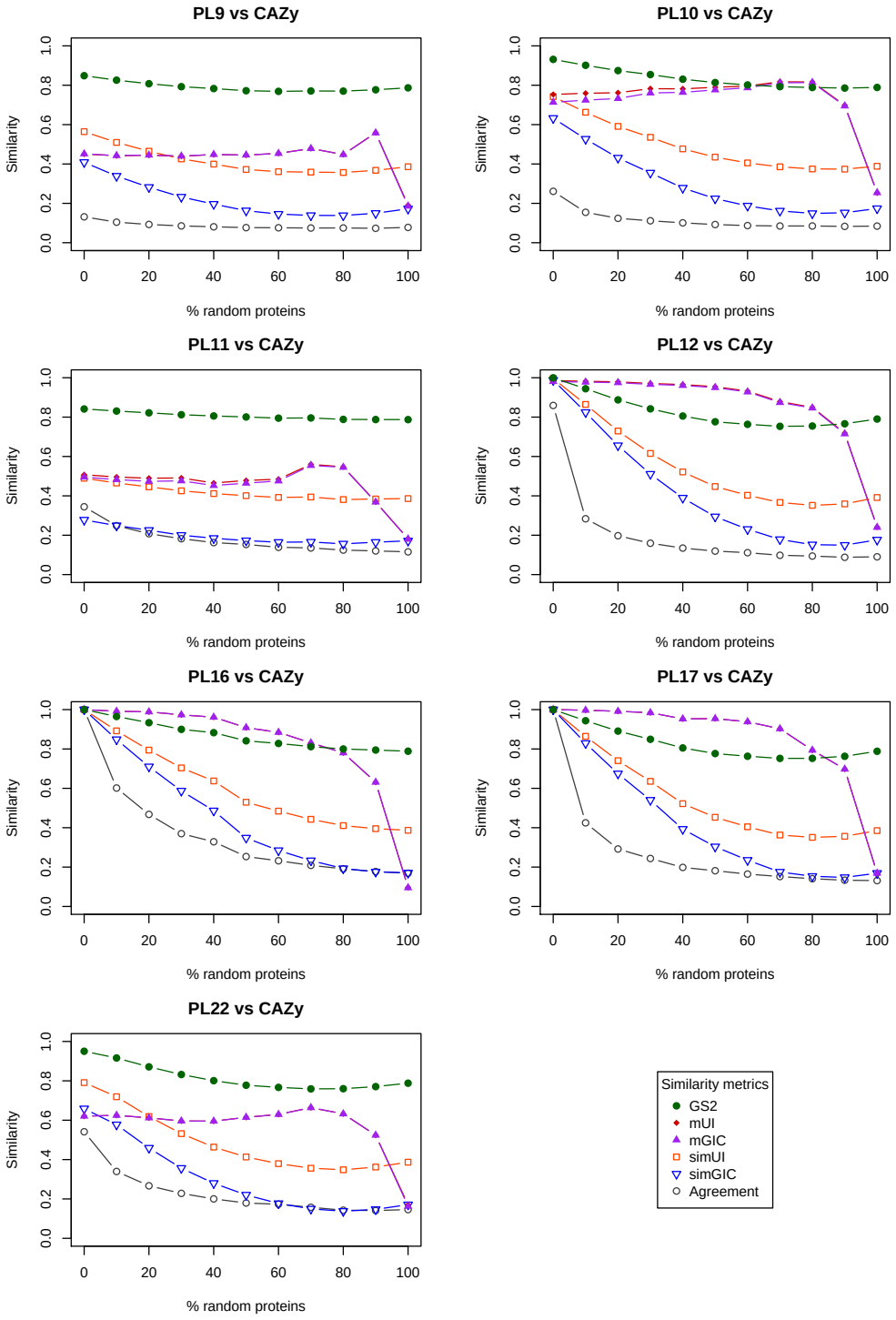
Coherence assays

From the analysis of Figure 4.9 it can be seen, as expected, that the similarity reported by each metric generally decreases as noise (random proteins) are increasingly added (and replace the original proteins) in each of the tested PL families. The Agreement metric is the least noise resilient metric, as can be seen by both the generally low values it reports and the steep declines after adding small amounts of noise to family sets with previously high agreement. This property is most evident in mono-functional families like PL5, PL16 and PL17 and also PL12 where the introduction of 10% random proteins produces a sharp decline in the reported values. This occurs because this naive metric only equates the average of annotation term frequencies in each protein family (or set). This metric was chosen and used as the overall baseline.

The simUI and its derivative simGIC, as expected, have a similar behaviour because simGIC is a IC-weighted version of simUI. Furthermore, in the obtained results (Appendix A) it is noticeable that simGIC presents a greater resolution than simUI (average range of 0.57 against a range of 0.46, respectively, as can be computed from Table 4.4), a behaviour that was also previously reported by Pesquita *et al.* (2009) in their assessment of semantic similarity metrics. In contrast, the GS^2 metric has the smallest resolution (for the tested sets) of all the tested metrics showing an average range of 0.18. In addition, to offering a smaller range of values (and a thus lower resolution) it is important to notice that reported values for this metric fall within the 0.75-1.0 range of similarity. Given that it is expected for protein (enzyme) families to have functionally similar proteins it would also be expected (and optimal) that these families would display higher coherence. However, when the unadulterated families are considered some of them do not provide the necessary annotations supporting such high global set functional coherence values, especially when considering values produced from the sets 100% randomized. This point will be revisited and discussed further ahead in more detail.



(a) *Similarity results for families PL1-PL8*



(b) *Similarity results for families PL9-PL12, PL16, PL17, PL22*

Figure 4.9: *Plots of the average similarity as measured by six different metrics, for 15 PL protein families (from the CAZy database) and their derived sets. These sets were made by replacing the original proteins with increasing amounts (of 10% increments; 100 iterations) of random proteins (taken from the CAZy database).*

metrics	PL1	PL2	PL3	PL4	PL5
Agreement	0.122	0.391	0.260	0.368	0.874
simUI	0.298	0.497	0.620	0.376	0.650
simGIC	0.356	0.539	0.825	0.458	0.853
mUI	0.139	0.214	0.671	0.353	0.801
mGIC	0.147	0.216	0.672	0.343	0.802
GS^2	0.137	0.224	0.238	0.177	0.246

metrics	PL6	PL7	PL8	PL9	PL10
Agreement	0.405	0.201	0.180	0.058	0.178
simUI	0.386	0.432	0.548	0.207	0.368
simGIC	0.429	0.542	0.660	0.27 0	0.484
mUI	0.469	0.501	0.329	0.368	0.564
mGIC	0.474	0.505	0.285	0.372	0.559
GS^2	0.175	0.129	0.224	0.080	0.146

metrics	PL11	PL12	PL16	PL17	PL22
Agreement	0.229	0.771	0.831	0.869	0.400
simUI	0.108	0.644	0.613	0.649	0.443
simGIC	0.122	0.838	0.829	0.853	0.521
mUI	0.378	0.744	0.903	0.831	0.494
mGIC	0.373	0.741	0.905	0.831	0.501
GS^2	0.054	0.247	0.211	0.248	0.191

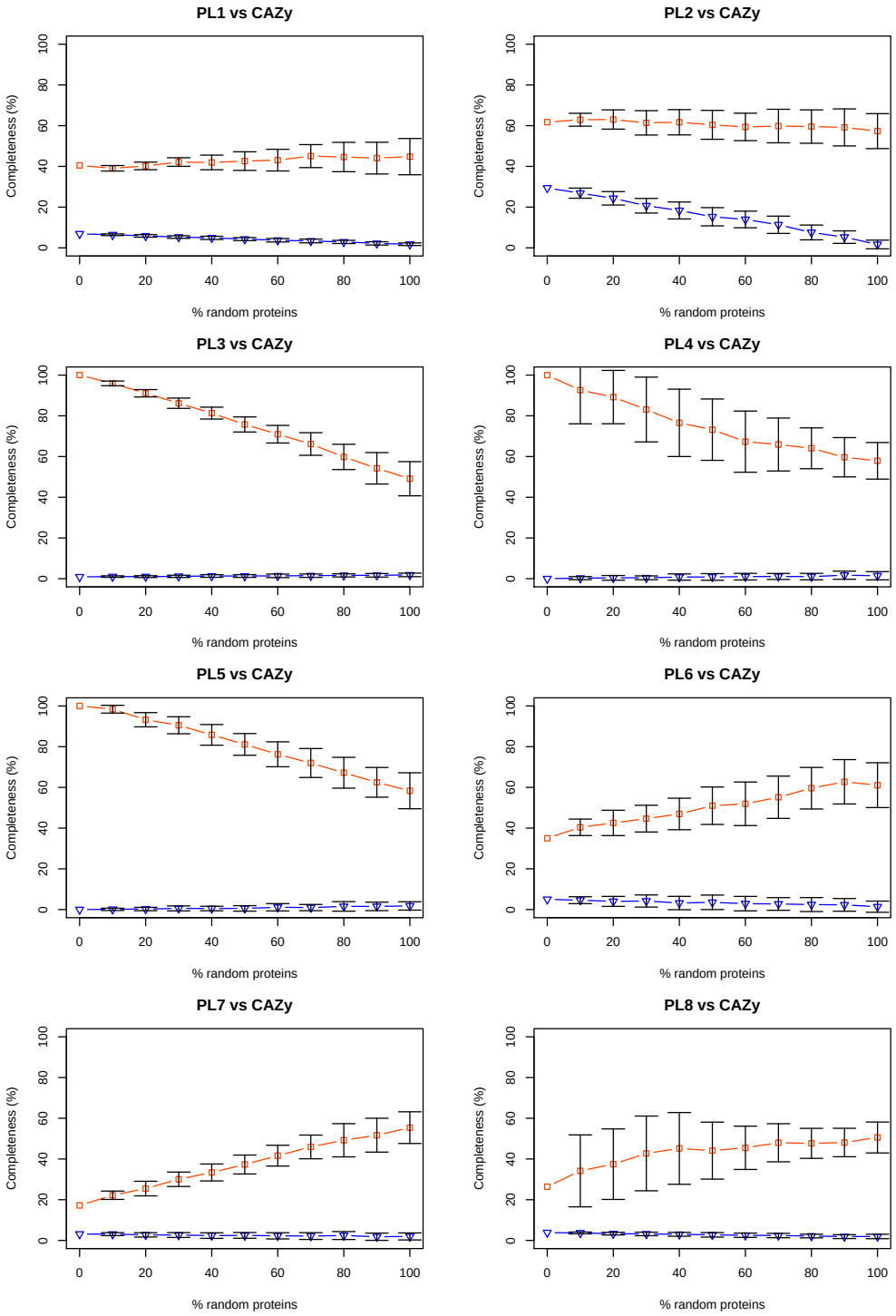
Table 4.4: *Difference between maximum and minimum values reported for each tested metric (Agreement, simUI, simGIC, mUI, mGIC, GS^2) against each PL family and iterations of derived respective sets created by insertion of increasing amounts of random proteins (from CAZy) into the original families.*

The mUI and mGIC (such as the metrics they are derived from) also display, as expected, similar behaviours to each other. In fact, for most of the tested PL families and their respective degenerate sets the reported values are very similar. However, unlike the other tested metrics mUI and mGIC are very resilient to noise (replacement with random proteins). That is evident from the gradual curves in Figure 4.9 that in most families plateau until higher levels of randomization and typically only fall abruptly after 90% random proteins replacing the original proteins in a given family. This resilience to noise is conferred by the term enrichment step which pre-selects only the subset of proteins that are annotated with the terms found to be statistically significant by the enrichment procedure. Thus, this is an important factor to consider when analysing the results provided by these two metrics. They were engineered to capture local (subset) functional coherence so for a comprehensive evaluation they should only be used in an analysis that also simultaneously considers the annotation coverage within the analysed set. This also explains the observed peaks at high noise levels in some of the families (PL2, PL6, PL9, PL11) where a small number of terms annotated a small subset of proteins and thus create a local similarity effect. However, for this work this behaviour is advantageous because the underlying assumption is that each protein family shares core annotations that define the group role of that set of proteins. Thus, by using a term enrichment technique the purpose is to target and select these core annotation terms in order to provide potential seeds for annotation extension within that set.

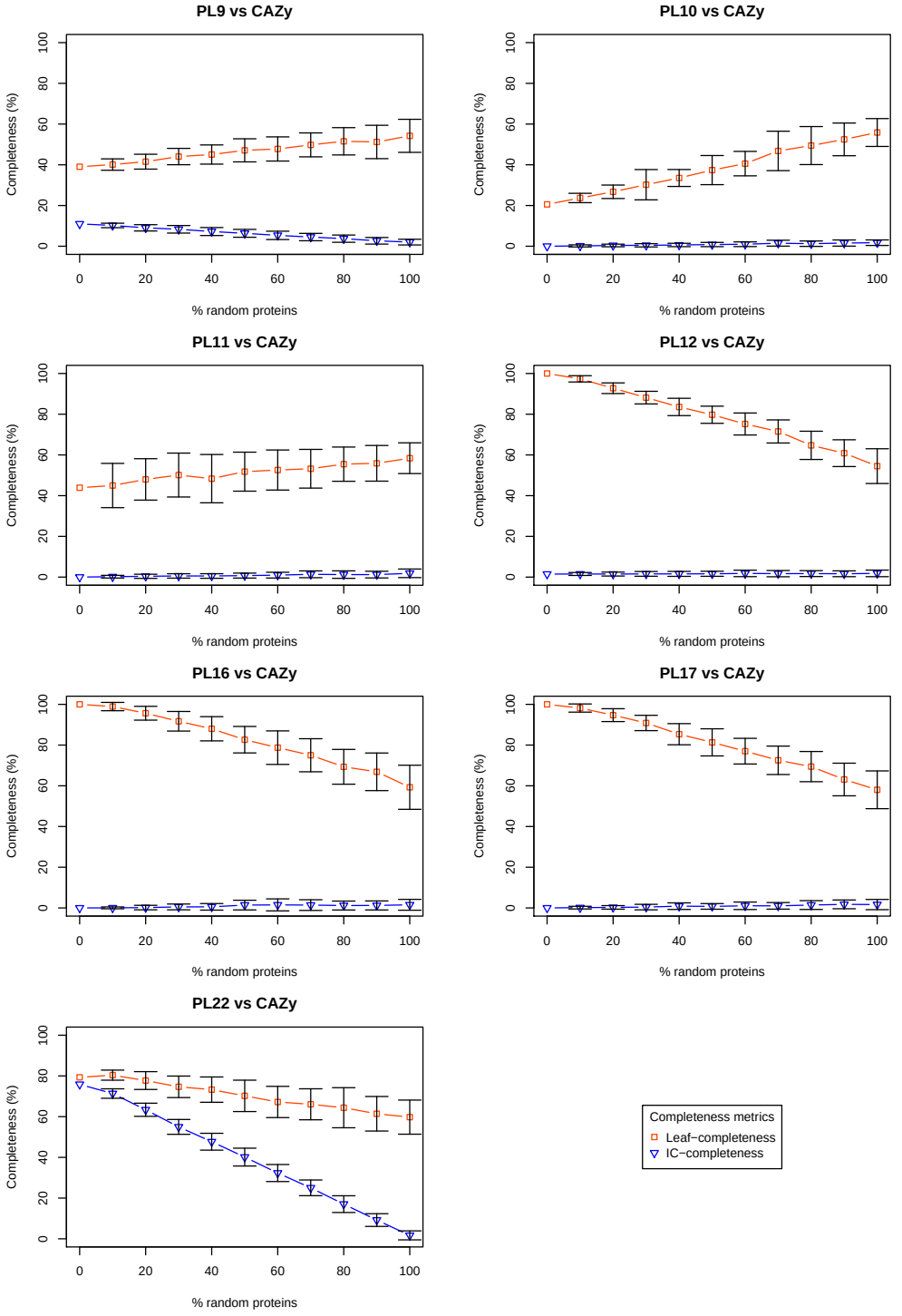
Completeness assays

The two developed completeness metrics were also applied to the protein sets created during the randomization resilience tests previously described. These metrics are both naive metrics. Considering the Leaf-completeness metrics and analysing the plots in Figure 4.10 the only observable trend is that with increasing set randomization the degree of Leaf-completeness converges upon a range around 50%-60% completeness.

This is actually related to the available CAZy annotation corpus and distribution of its annotation terms. Therefore, this metric serves only as broad indicator



(a) *Completeness results for families PL1-PL8*



(b) *Completeness results for families PL9-PL12, PL16, PL17, PL22*

Figure 4.10: *Plots of the average completeness as measured by the Leaf-completeness and the IC-completeness (with respective standard deviations), for 15 PL protein families (from the CAZy database) and their derived sets. These sets were made by replacing the original proteins with increasing amounts (of 10% increments; 100 iterations) of random proteins (taken from the CAZy database).*

of the distribution of annotation terms for any given protein set. Regardless, it can be used as a quick cue indicating potential incompleteness in a set (under the assumption that any protein in a set must at least be annotated with one leaf-term).

Regarding the IC-completeness metric its assumption is similar to the previous completeness metric. However, whereas in Leaf-completeness each protein needs at least one leaf-term to contribute to the protein set completeness, in the IC-completeness each protein needs at least one annotation term with an IC score above a chosen threshold. In this work, the chosen threshold was 0.7 which corresponds to the 50th percentile of the annotation corpus and it was chosen this permissively in order to be more inclusive. However, in Figure 4.10 low IC-completeness values can be seen, even for most of the original families (before any noise introduction). Even well annotated mono-functional families like PL16 and PL17, where a single annotation term (*hyalurononglucosaminidase activity* and *poly(beta-D-mannuronate) lyase activity*, respectively) conveys a specific function did not report expectable values for IC-completeness. This stems from one of the limitations of using IC as an indicator of term specificity: there are some terms, that despite describing very specific functions, become very frequent in annotation corpora. This happens because these functions are biologically important across several species which consequently leads those terms to have lower IC scores. Therefore, the IC-completeness metric should not be used without analysing each set and a unique threshold per set might be more suitable. Regardless, this IC-completeness as it was performed with a threshold of 0.7 (50th percentile) allowed to characterize the PL collection as lacking in term specificity for most of its families.

4.4 Protein Annotation Extension

The previously presented studies and assays of the current protein annotation panorama led to the identification of several deficiencies regarding coverage and specificity. Based on those studies a modular framework is proposed here for the extension of protein annotation within protein (enzyme) families (or other

functionally similar sets). In this section the proposed framework of this thesis is presented along with the implementation of the annotation extension module.

4.4.1 Methods

In this thesis, the used protein families (from specialized databases) are assumed to be functionally coherent (to a greater or lesser degree) even if their annotation does not always reflect this (due to under-annotation). That underlying assumption is used because the methodologies typically employed to aggregate the proteins into families rely on techniques used to find similarities between sequences and the identification of common domains or functional modules. Figure 4.11 illustrates the proposed framework for annotation extension in under-annotated protein families.

In the proposed framework, for each family, significant terms are discovered through statistical enrichment techniques. The aim is to find the most specific terms that may be under-annotated in a family and then use the proteins annotated to these significant terms as seeds for potentiating the annotation expansion for other proteins annotated only up to parent terms of these enriched terms. The framework is modular and for this thesis, as the statistical enrichment techniques a combination of the Fisher Exact test and the Elim method was used as previously described in Chapter 4. In addition, visualization was also used to iteratively assist in choosing target terms from the annotation graphs of each given family that could potentially be extended into more generically annotated proteins on each respective family. For the visualization of annotation graphs and respective term enrichment analysis of the protein families, GRYFUN, a specific web application was developed. This web application will be presented and further discussed in Chapter 5. The proteins annotated with the selected terms for the potential annotation extension in each family (set) were used to automatically generate multiple sequence alignments (MSA) resorting to the MAFFT (Katoh & Toh, 2008) program (using its default settings). These MSA were subsequently used to build profiles with the HMMER (Finn *et al.*, 2011) program (also using the default settings). For each term (previously found significant) on which extension is going to be attempted, the target proteins in a family submitted to

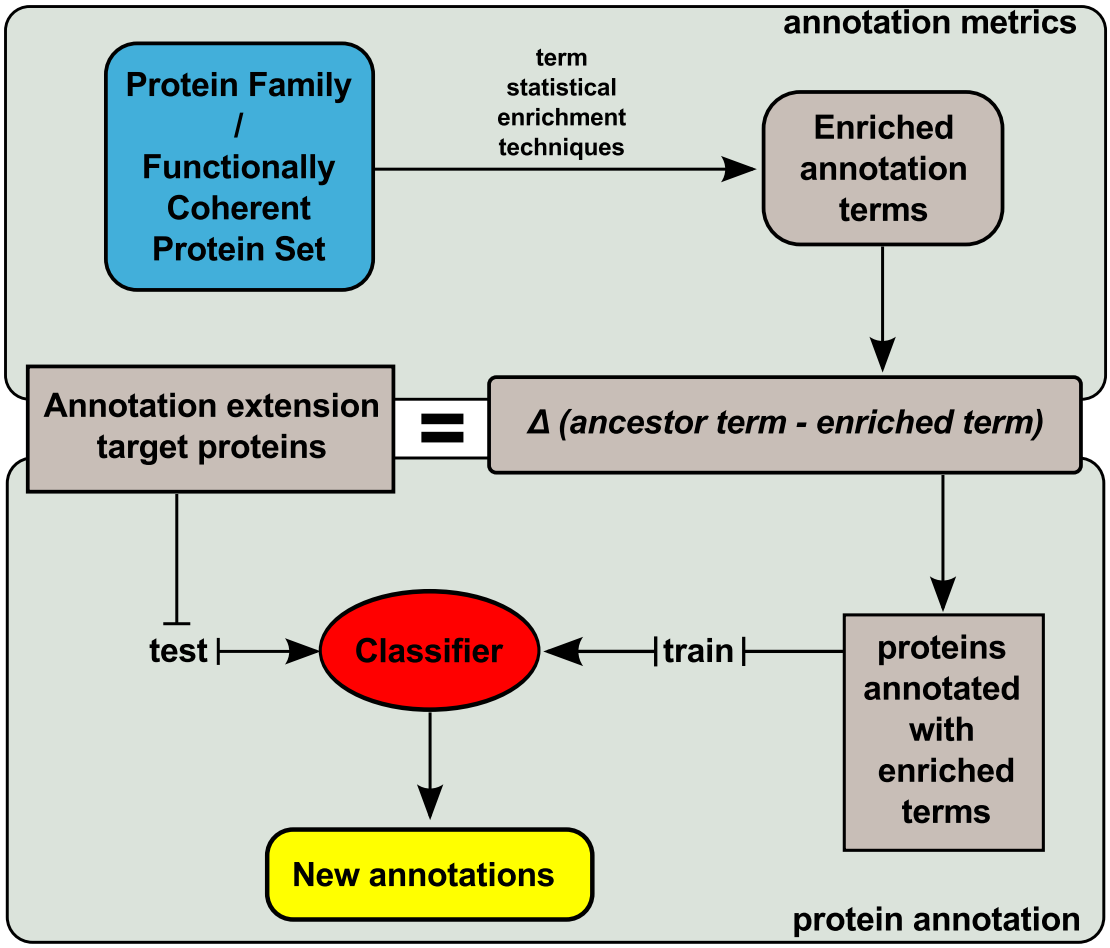


Figure 4.11: *Proposed framework for the measuring of annotation coherence and extension of annotation in under-annotated proteins families.*

the proposed annotation extension are the proteins annotated only as far as the parents of that term.

The testing and validation of this annotation extension module was through a repeated random sub-sampling validation, using a testing split of 33% over 10 iterations and was performed on a set of enriched terms from selected PL families. For each set of "training" proteins (67% splits) an HMM profile was created and each of those proteins was then used for a self-scan against the created profile. The average of all the resulting best domain scores hits was calculated and a

decision threshold was created by subtracting their respective standard deviations as shown in Equation 4.1.

$$Decision\ threshold[t] = \frac{\sum hit\ score_{protein}}{number\ of\ hits} - \sigma, \quad \forall\ proteins\ \in annotations[t] \tag{4.1}$$

The test sets (33% splits) were then scanned against the HMM profiles and proteins scoring above the classification threshold (as shown in Equation 4.1) were recognized as recipients for annotation extension. Considering that in this assay the assumption is that of an open world, where a protein not being selected does not necessarily exclude it from having the term annotation it was tested against, only the average recall (fraction of retrieved positively annotated proteins) values were calculated for the classification results of the 10 iterations of the random repeat sub-sampling validation.

The MAFFT and HMMER programs were chosen due to their ready availability and ease of interfacing. In order to use them according to the above described proposed framework a custom Python wrapper library to interface with their binaries was newly written.

4.4.2 Results

The implementation of the previously described repeated random sub-sampling validation technique was applied to proteins annotated with terms found significant in PL families under study and yielded the average recall results presented (with respective standard deviation) in Table 4.5. These results are also plotted in Figure 4.12 superimposed on the barplot indicating how many sequences each of the term extension repeated random sub-sampling validation assays used to build their MSA/HMM profile.

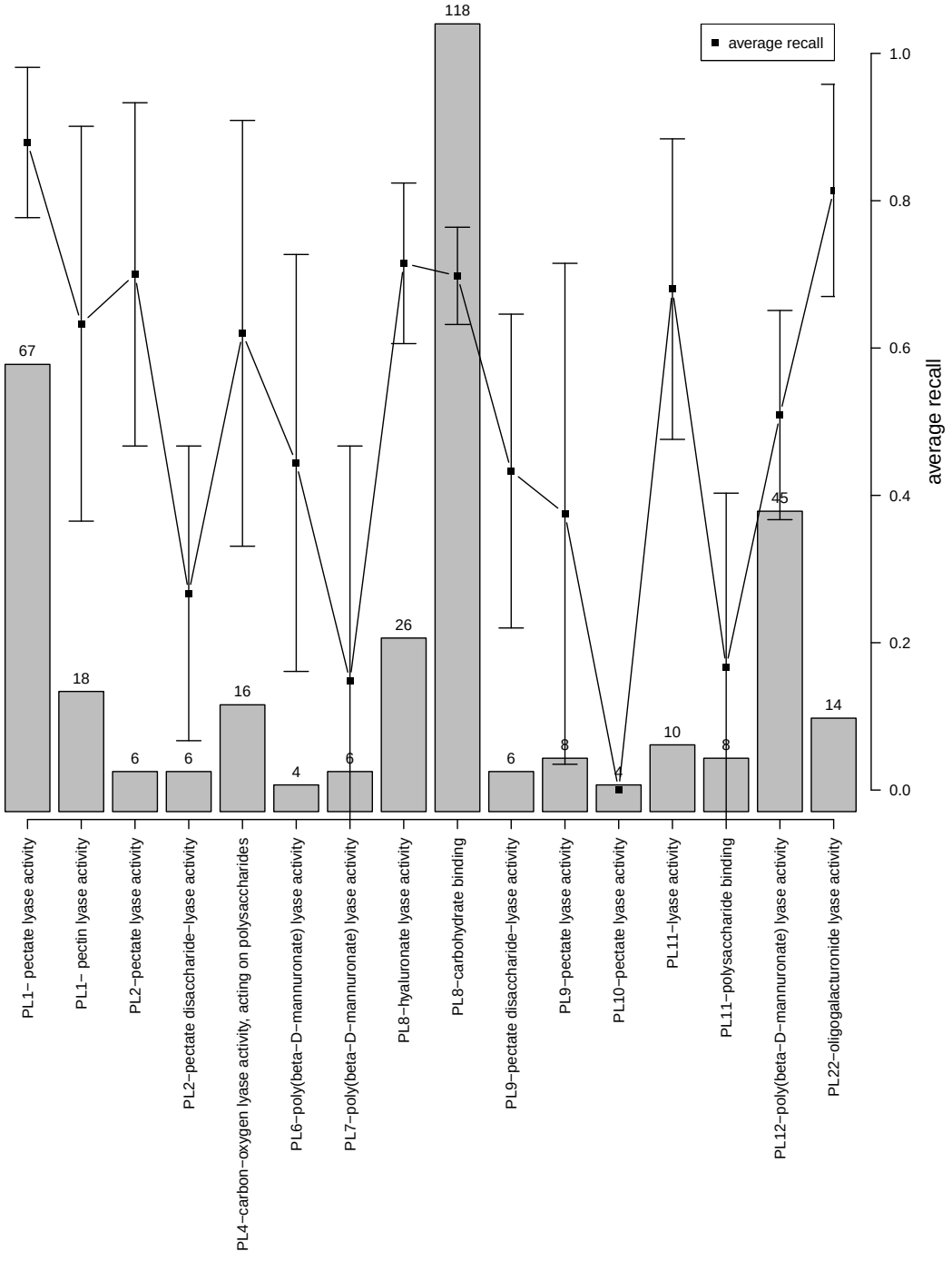


Figure 4.12: *Plot of the average recall obtained for each extension term and their respective standard deviation. Additionally, the number of sequences used to build each MSA/HMM is superimposed as a barplot.*

Family	term name	recall	σ
PL1	pectate lyase activity	0.879	0.102
PL1	pectin lyase activity	0.633	0.268
PL2	pectate lyase activity	0.700	0.233
PL2	pectate disaccharide-lyase activity	0.267	0.200
PL4	carbon-oxygen lyase activity, acting on polysaccharides	0.620	0.289
PL6	poly(beta-D-mannuronate) lyase activity	0.444	0.283
PL7	poly(beta-D-mannuronate) lyase activity	0.148	0.319
PL8	hyaluronate lyase activity	0.715	0.109
PL8	carbohydrate binding	0.698	0.066
PL9	pectate disaccharide-lyase activity	0.433	0.213
PL9	pectate lyase activity	0.375	0.340
PL10	pectate lyase activity	0.000	—
PL11	lyase activity	0.680	0.204
PL11	polysaccharide binding	0.167	0.236
PL12	poly(beta-D-mannuronate) lyase activity	0.509	0.142
PL22	oligogalacturonide lyase activity	0.814	0.144

Table 4.5: *Recall and respective standard deviations for the repeated random sub-sampling validation procedure on chosen significant terms on a set of tested PL families.*

4.4.3 Discussion

The selected PL families for the repeated random sub-sampling validation were PL1, PL2, PL4, PL6 to PL12 and PL22. The reason for this selection was that other PL families presented no detectable annotation extension opportunities (PL3, PL5, PL16 and PL17). As before, families PL13, PL14, PL15, PL20 and PL21 were not considered due to their low number of annotations.

As can be seen both in Table 4.5 and in Figure 4.12 the recall values resulting from the assays vary considerably. The first thing that must be taken into account is that the assays were run using the default settings for both MAFFT and HMMER. The tweaking of the settings and parameters or the manual adjustment of the automatically generated MSA by a domain expert would most likely increase the obtained recall values. However, the focus of this thesis was not on optimizing this part of the framework and instead provide a baseline upon which anyone can improve (hence the proposed modularity of the framework). Regarding, the

specific assayed term annotation extensions, the aforementioned validation done on the families PL1, PL4, PL8, PL12 and PL12 significant terms (with the addition of *pectate lyase activity* for family PL2 and *lyase activity* for family PL11) yielded reasonably high recalls. These results in the most part are supported by the MSA/HMM profiles that were generated from higher numbers of protein sequences. The notable exceptions are the family PL2 pectate lyase activity and lyase activity in family PL11, which display reasonably high recalls despite having used smaller number of proteins to generate their MSA/HMM profiles. On the hand, the standard deviations are quite high for these, and most of the assays in this validation. Therefore, only the PL1-*pectate lyase activity*, PL8-*hyaluronate lyase activity*, PL8-*carbohydrate binding*, PL12-*poly (beta-D-mannuronate) lyase activity* and PL22-*oligongalacturonide lyase activity* offer a relatively reasonable confidence ($\sigma < 0.2$). This simple assay shows that a reasonable number of protein sequences is needed to create a MSA/HMM profile seed for reliable annotation extension. Further validation of the annotation module (within the complete context of the proposed framework) will be presented in Chapter 6.

Summary

In this chapter an exploratory analysis of the CAZy database annotation space was described. Additionally, metrics for annotation completeness, agreement and coherence were introduced and their resilience assayed. Furthermore, the module for annotation extension in protein families of the proposed framework was presented and the performance of its implementation was measured for its recall.

The next chapter describes GRYFUN, the implementation of the functional coherence analysis module of the proposed framework, as a web application.

Chapter 5

GRYFUN

The coupling of enrichment analysis with protein annotation visualization can improve the analysis because it enables the identification of existing relationships between annotation terms found to be enriched. Bioinformatic tools like GOBar (Lee *et al.*, 2005), GOLEM (Sealfon *et al.*, 2006), GOrilla (Eden *et al.*, 2009), StRAnGER (Chatziioannou & Moulos, 2011) among several other tools provide this combination of enrichment analysis and annotation visualization.

Consequently, these tools produce visual representations of the graph structures that subsume the terms annotating a target protein (or gene) set in addition to the enrichment values, that in some cases are also incorporated in the graph visualization (for e.g. as color gradients). However, instead of using one of these currently available tools a specific tool named GRYFUN (GRaph AnalYzer of FUNctional annotation) was developed to meet specific demands. GRYFUN, despite its similarities with the aforementioned applications is particularly designed and focused on the analysis of the functional annotations in protein families (or in functionally related sets of proteins) regarding their annotation coherence, cohesiveness and extension potential (Bastos *et al.*, 2015).

5.1 Implementation and Input

GRYFUN is a web-based application developed on the web2py Web Framework (`http://www.web2py.com/`) which relies on the Python programming language. This framework was chosen because it focuses on rapid development and

follows a Model-View-Controller design making the application easy to extend. It is also easy to run and deploy and offers protection against the most common security issues. Furthermore, the graph drawing component of this application was developed around the GraphViz (Ellson *et al.*, 2001; Gansner & North, 2000) visualization software package. Graphviz is a well known open source graph visualization software and is often used in bioinformatic applications. Interactive elements of this application are mostly handled by customized Javascript code. GRYFUN is available and can be publicly accessed at: `http://xldb.di.fc.ul.pt/gryfun/`. Additionally, all the GRYFUN code was made available as open source under an MIT license and deposited in a GIT server at: `https://bitbucket.org/hpbastos/gryfunserver.git`.

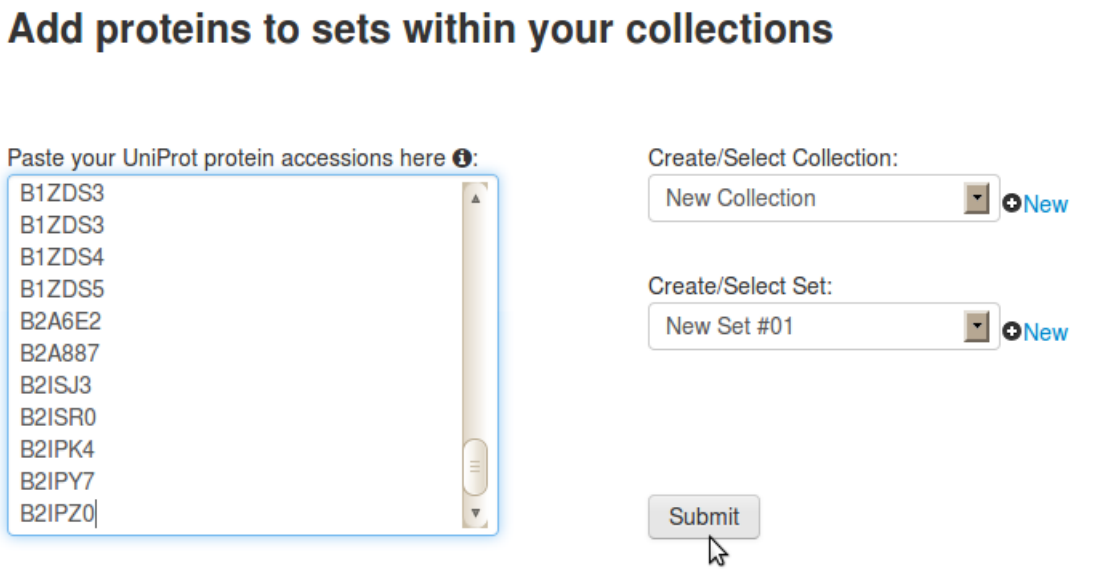


Figure 5.1: *GRYFUN protein Set input interface*

Currently, GRYFUN only accepts UniProt accession numbers as input protein identifiers, as depicted in Figure 5.1, in order to create a user protein Set. Input identifiers are validated against those present in a background database (UniProtKB release 2014-02). Only matched protein identifiers having at least one GO annotation are added to the input Set. The source of the GO annotation mapping used in GRYFUN is supplied by the GOA project (Barrell *et al.*, 2009).

In addition to enabling the creation of Sets and Collections, GRYFUN also enables each user the individual deletion of any of their Sets and Collections.

Each Set must belong to a Collection, which besides providing a way to group Sets that share some functional similarity, can also and consequently create a coarser level of granularity. Most importantly, a proper use of the Collection/Set organization is paramount for computing meaningful GO term enrichment p-values. For any given Set, the statistical tests are applied on the remaining Sets in that Collection as the background set to determine the statistical significance of the enrichment of any given annotation term in the Set being explored.

The input proteins in each Set are expected to have a close degree of functional similarity, such as is the case of functional protein families or other groups of functionally related proteins. Alternatively, a Set can host dissimilar proteins if the intended purpose is just to navigate the generated annotation graph and manually sort and select sub-sets of proteins.

5.2 Graph Visualization and Interaction

GRYFUN enables the generation of the annotation graph for any protein Set (within a given user Collection) under the context of each one of the three GO orthogonal ontologies (biological process, molecular function and cellular component). That functionality can be accessed through the Explore page on this web application. Figure 5.2 depicts the selection of the *molecular_function* ontology aspect for generating the GO annotation graph for a Set named PL1 (corresponding to the family with the same name) of the *Polysaccharide Lyase* Collection. In the depicted query, all Evidence Code are considered (default) but each user has the ability to filter the annotations with only the evidence codes that are relevant for their own work.

The annotation graphs generated by GRYFUN are similar (and dependent) on GO graphs, however they present with a couple of important differences. A GO graph is meant to denote relationships between terms, so while each term

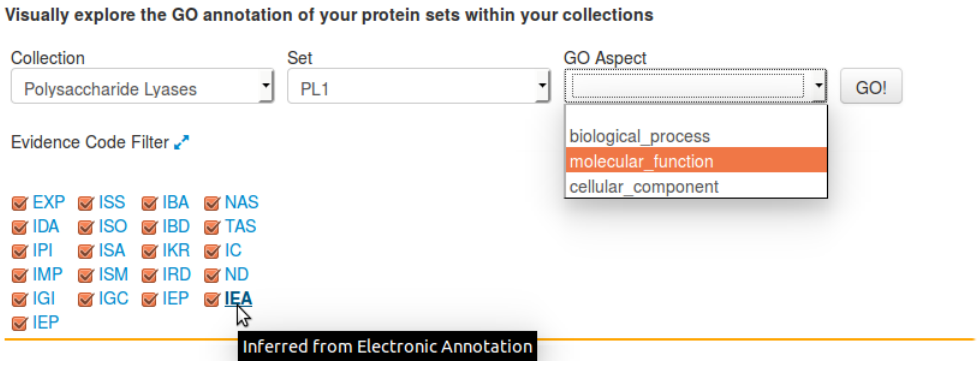


Figure 5.2: *GRYFUN's graph generation menu from its Explore Page.*

is represented by a node the relations in-between them are represented by graph edges. Figure 2.7 shows a GO sub-graph depicting nodes of the *biological process* GO aspect connected by *is_a* edges. Each of these edges starts at child nodes (terms) and point towards parental nodes (terms), and thus denote the existing hierarchical relationships between terms. Additionally, all terms converge into a common root node, thus leading to the true path rule that states that "*the pathway from a child term all the way up to its top-level parent(s) must always be true*" (Gene Ontology Consortium, 2000).

On the other hand, in the annotation graphs, like the one shown in Figure 5.3 the edge direction is reversed. Every protein in a Set leading to an annotation graph is mandatorily annotated to at least the root term (*biological_process* in this case). Depending on how well annotated any given protein is, it will "flow down" the graph towards more specific nodes. That "flow" can be immediately discernible from the annotation graph given that the edge thickness is generated in proportion to the number of proteins that "flow down" from one parent node to its child node. Therefore, by representing the "annotation flow" on the graph image, an immediate visual cue is provided regarding the annotation terms that are more represented in any given protein Set.

Hovering the mouse cursor over any graph node will reveal the associated term and its annotation frequency within the current Set as a tooltip. On the other hand, clicking on any of the nodes (white nodes: inherited annotations,

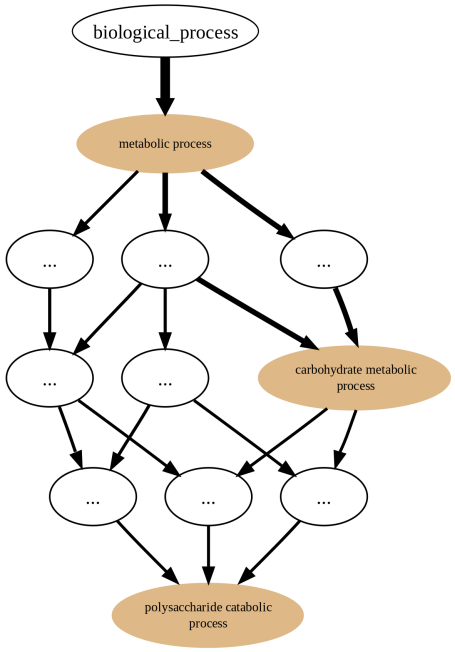


Figure 5.3: *Example annotation graph of a sample protein set for the GO biological_process aspect.*

color nodes: direct annotations) will dynamically generate a floating window containing a list of the respective UniProt accession numbers annotated to that term within the Set, as shown in Figure 5.4. Furthermore, those floating windows also display the respective species names, alphabetically sortable and segregated into Superkingdoms. These lists can be exported into plain TSV files. Also, any number of floating windows, up to the number of nodes in the currently displayed graph, can be open simultaneously. Furthermore, these windows can be dragged anywhere on screen and collapsed and expanded as required. Most importantly, these floating windows enable access to one of the most interesting features in GRYFUN, graph *re-rooting*. This feature is similar to the GOLEM (Sealfon *et al.*, 2006) *focus* feature which reduces the graph to a selected GO annotation term and its vicinity (parents and children). On the other hand, the *re-root* feature in GRYFUN allows the selection of any non-leaf term node in the annotation graph and the generation of a new sub-graph rooted at the term represented by the chosen node. After this *re-rooting* operation, and despite the Set remaining whole, only the proteins annotated with the new temporary

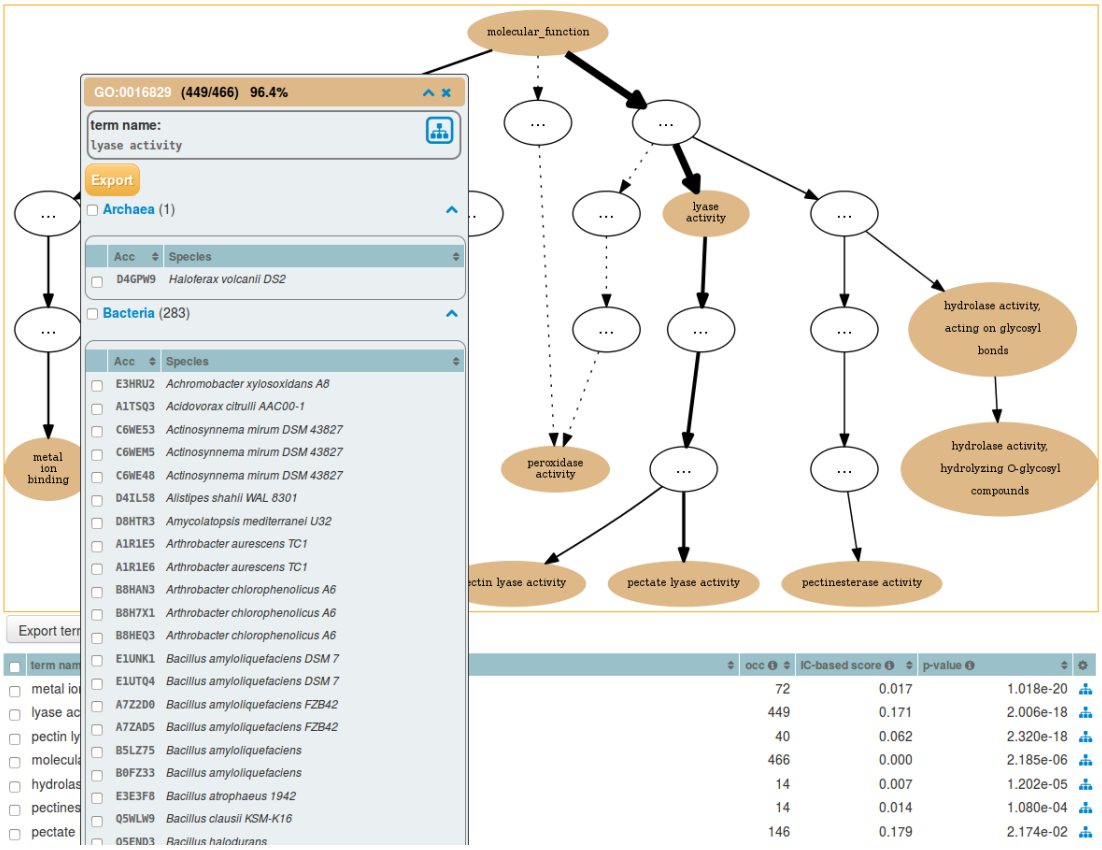


Figure 5.4: *Partial display of the Explore page following a graph generation with the node (for term lyase activity) floating information window displayed on screen as well.*

chosen root are considered during the generation of a new annotation sub-graph that subsumes all of their annotations that are children of the new chosen term. Hence, this feature enables the focus on more specific functional branches and terms of interest while abstracting from terms that sometimes describe accessory activities that despite being associated to some proteins in a given Set can be considered to be noise.

5.3 Supporting information and statistics

The actual generated annotation graph at the Explore page is preceded by an header with general information pertaining to the Set currently being inspected. On the other hand, a table list succeeds the graph and contains metrics associated

to the terms that annotate the current Set.

5.3.1 Explore page: Header

The Explore page in GRYFUN, atop of each dynamically generated annotation graph, also displays an header with the names of the Collection and Set currently being queried as depicted in Figure 5.5. Additionally, the header displays:

- *Total set size*: number of total UniProt entries on the current Set.
- *Annotation coverage*: percentage (and number) of UniProt entries of the current Set annotated in the current GO aspect.
- *Current root*: GRYFUN allows to root the annotation graph at different terms other than the actual GO aspect root terms.
- *Current coverage*: when a new root is chosen it shows how many UniProt entries are annotated to that new root term.

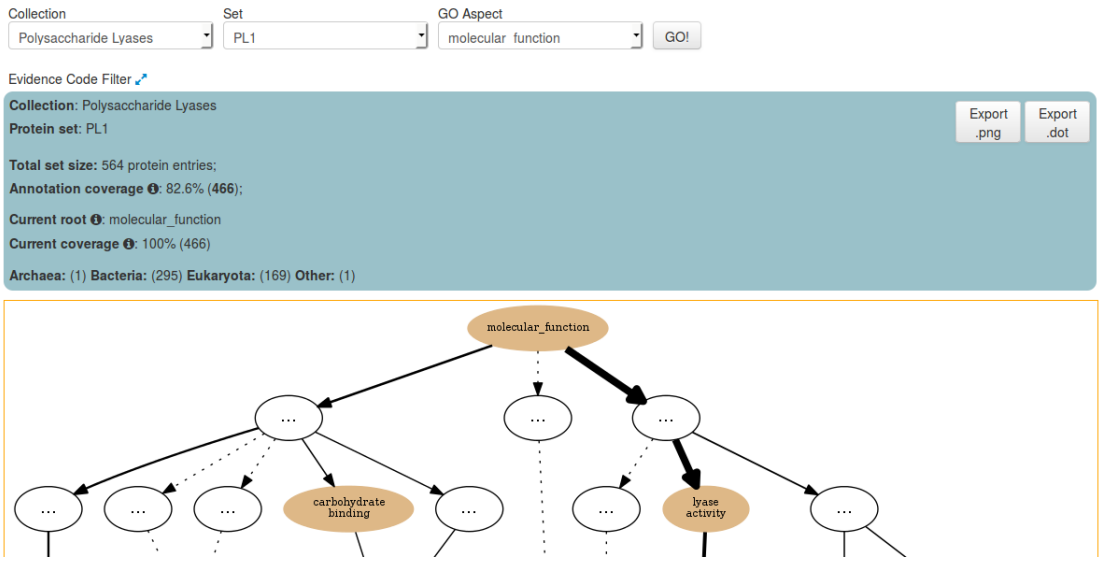


Figure 5.5: *Header of GRYFUN's Explore page following an annotation graph generation.*

Additionally, the last line of the header shows the breakdown of the UniProt entries in the current Set regarding each taxonomical Superkindgom. Furthermore, the header contains two convenience buttons: *Export .png* and *Export .dot.* that enable opening a new tab/save the image file of the generated annotation graph and the underlying graph file (in dot format) respectively, so that a user might use these with external applications.

5.3.2 Explore page: Footer

The last section of an Explore resulting page is the terms information table such as the one show in Figure 5.6. Similarly to the tables within the floating windows this term table is also sortable and the selected entries (alongside with their associated data) can be exported as a TSV file with just a press of a button (Export terms). Besides, each annotation term name, these tables display the following fields:

- *occ*: number of term occurrences, that is, how many proteins are annotated with that term in the given Set.
- *IC-based score*: IC-based score, relying on GOA as the annotation corpus.
- *p-value*: nominal p-value for the term enrichment (via Elim procedure) in the current Set relative to the remainder of the Collection (background).

The commonly used term-for-term approach is applied in GRYFUN to detect GO term annotation enrichment in protein Sets. The p-values are calculated using a Python implementation (`https://pypi.python.org/pypi/fisher/`) of the Fisher's exact test. Furthermore, the Elim strategy was implemented using the Python programming language (and using a significance level of 0.05) to correct the p-values in order to mitigate the propagation issue derived from the GO graph nature and usage of the term-for-term approach, as previously described.

Export terms

term names	occ	IC-based score	p-value
metal ion binding	72	0.017	1.018e-20
lyase activity	449	0.171	2.006e-18
pectin lyase activity	40	0.062	2.320e-18
molecular_function	466	0.000	2.185e-06
hydrolase activity, acting on glycosyl bonds	14	0.007	1.202e-05
pectinesterase activity	14	0.014	1.080e-04
pectate lyase activity	146	0.179	2.174e-02
heme binding	1	0.000	3.024e-01
peroxidase activity	1	0.001	3.024e-01
hydrolase activity, hydrolyzing O-glycosyl compounds	13	0.007	5.187e-01
cellulose binding	2	0.002	5.854e-01
polysaccharide binding	6	0.006	7.412e-01
catalytic activity	463	0.022	1.000e+00
carbon-oxygen lyase activity, acting on polysaccharides	185	0.170	1.000e+00
oxidoreductase activity, acting on peroxide as acceptor	1	0.001	1.000e+00
carbon-oxygen lyase activity	185	0.098	1.000e+00
antioxidant activity	1	0.001	1.000e+00
ion binding	72	0.010	1.000e+00
cation binding	72	0.016	1.000e+00
tetrapyrrole binding	1	0.000	1.000e+00
binding	87	0.006	1.000e+00
carbohydrate binding	14	0.008	1.000e+00
organic cyclic compound binding	1	0.000	1.000e+00
pattern binding	6	0.006	1.000e+00
hydrolase activity	28	0.005	1.000e+00
hydrolase activity, acting on ester bonds	14	0.005	1.000e+00
heterocyclic compound binding	1	0.000	1.000e+00
carboxylic ester hydrolase activity	14	0.009	1.000e+00
oxidoreductase activity	1	0.000	1.000e+00

Figure 5.6: *Footer table of term names and respective metrics in GRYFUN's Explore page following the generation of an annotation graph.*

5.4 GRYFUN usage examples

As an example to showcase the use of GRYFUN the PL1 Set (family) was chosen (from the PL families dataset) and its annotation graph and associated statistics for the *molecular function* sub-ontology were generated. All evidence code annotation types were considered. The PL1 family/Set is comprised of 564 UniProt protein entries of which 466 are annotated with terms from the GO *molecular function* sub-ontology. This information is also displayed at the header of the generated page as can be seen in Figure 5.5. In addition, the header also displays information such as the Superkingdom taxonomical breakdown of the proteins in the current Set. As previously described, the central element of the dynamically generated page is an interactive annotation graph such as the one generated for the PL1 Set and depicted in Figure 5.7. Visual inspection of the graph immediately makes evident that the main *annotation flow* occurs

from the root term (*molecular function*) towards the two leaf-terms: *pectate lyase activity* and *pectin lyase activity*. Furthermore, by inspecting the path between the root term and these two leaf-terms, unsurprisingly, the term *lyase activity* can be found. Hence the graph confirms the expected dominant annotation with the term *lyase activity* and children sibling terms in a protein Set that is itself a subset of proteins belonging to Polysaccharide Lyase class protein family of the CAZy database classification. Hence, in the current PL1 Set example the *lyase activity* term node would be a good candidate for a *re-root* point. That is further supported by the p-values and IC-based score statistics (respectively $2,006 \times 10^{-18}$ and 0.171).

Figure 5.6 depicts the generated page footer containing the term names and respective statistics sorted by p-value. Given that the IC-based term score is the product of the IC of a term (in a given corpus) and its respective frequency in a given Set, it then provides a measure of *specific representativity* of a term in that Set. In other words, by having a high score the *lyase activity* term is one of the most frequent of the most specific annotation terms in the Set. However, since this is not a leaf-term there is a potential for annotation extension of the proteins not annotated beyond this term. Hence a GRYFUN *re-root* operation was performed on the *lyase activity* node yielding a new sub-graph as depicted in Figure 5.8. Thus, for this case three separate sets of proteins (one for each of the two leaf-siblings in the current Set and another for all the proteins annotated to the *lyase activity* term) can be exported and submitted to annotation analysis (manual or otherwise) that could lead to annotation extension as proposed previously in Chapter 4.

On the other hand, despite these terms deemed of interest and relevance being identified as enriched (statistically significant), in this example, the ranking of their p-values does not entirely match the *annotation flow*. However, the background against which the enrichment hypothesis was being tested was only the remainder of PL sets, and thus retaining an high degree of functional closeness, that is, a number of these activities would also be present in other Sets within this Collection. Nevertheless, when using all CAZy database families as the Collection (and hence statistical background), the enrichment results are closer to the

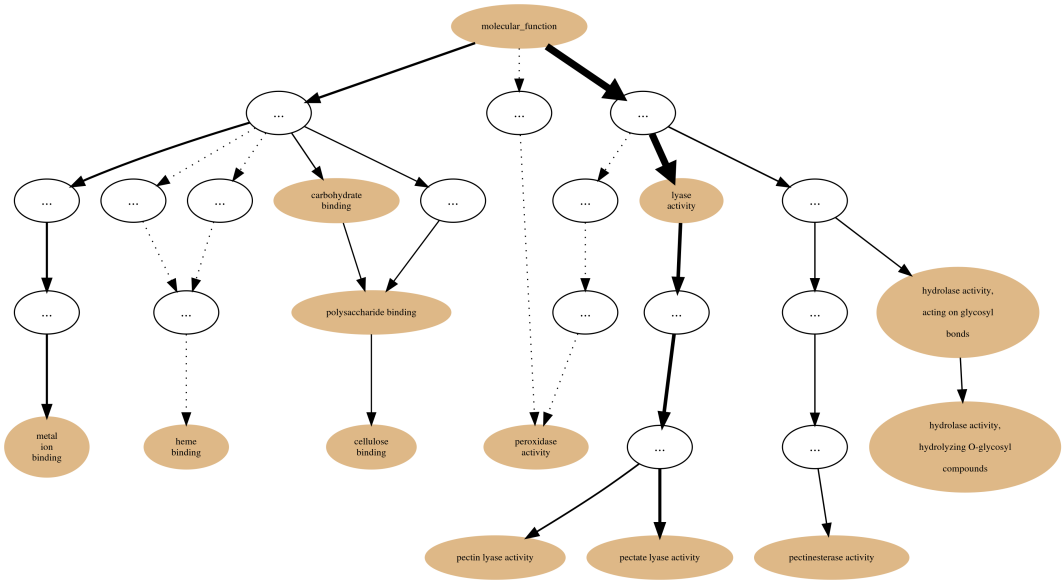


Figure 5.7: *Annotation graph subsuming the PL1 (CAZy family) Set GO molecular function sub-ontology annotations.*

expected values. Table 5.1 displays a sample of the term enrichment list, ranked by p-value, of the PL1 family (set) relation to a background of 237 CAZy families of catalytic classes Glycoside Hydrolases (GH), GlycosylTransferases (GT), Polysaccaride Lyases and Carbohydrate Esterases (CE). The top ranked terms here match the annotation flow as depicted in Figure 5.8, thus illustrating the importance of defining a good background in order to achieve a more accurate and reliable enrichment analysis.

In addition, the Evidence Code Filter can be used, for instance to filter out Inferred Electronic Annotations (IEA) and generate a new annotation graph for the PL1 Set containing only annotations typically regarded as being of higher quality, as previously discussed. The resulting annotation graph seen in Figure 5.9 is clearly simpler than the one in Figure 5.7 which was generated using all available annotations regardless of their Evidence Codes. Because the bulk of all annotations consist of IEA annotations the PL1 Set only has 32 out of 564 proteins with non-IEA annotations. Hence, this filtering focuses the PL1 Set on its annotations considered to be of higher quality but at the cost of coverage.

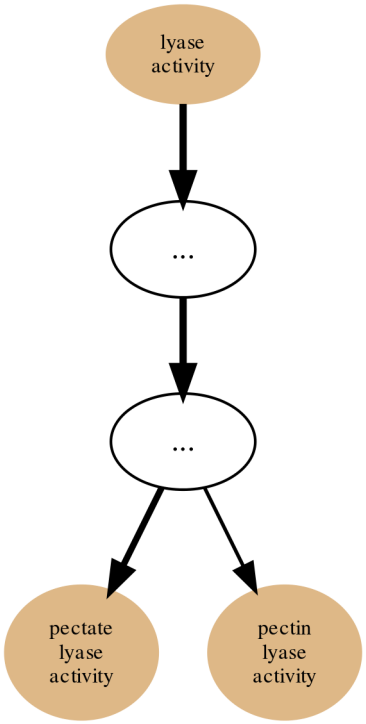


Figure 5.8: *Annotation graph of the PL1 (CAZy family) Set for the GO molecular function sub-ontology re-rooted at the lyase activity term.*

term name	p-value
lyase activity	$< 5.315 \times 10^{-248}$
pectate lyase activity	5.315×10^{-248}
pectin lyase activity	2.558×10^{-094}
metal ion binding	5.068×10^{-056}
molecular_function	6.276×10^{-008}
pectinesterase activity	1.146×10^{-007}
catalytic activity	7.547×10^{-004}
peroxidase activity	7.531×10^{-003}

Table 5.1: *Term enrichment p-values for the PL1 Set significant terms (alpha = 0.01) while using the complete CAZy Collection as background.*

Furthermore, the simplification of the graph also matches that of the previously shown term enrichment (using all annotations) thus reinforcing the previous enrichment results.

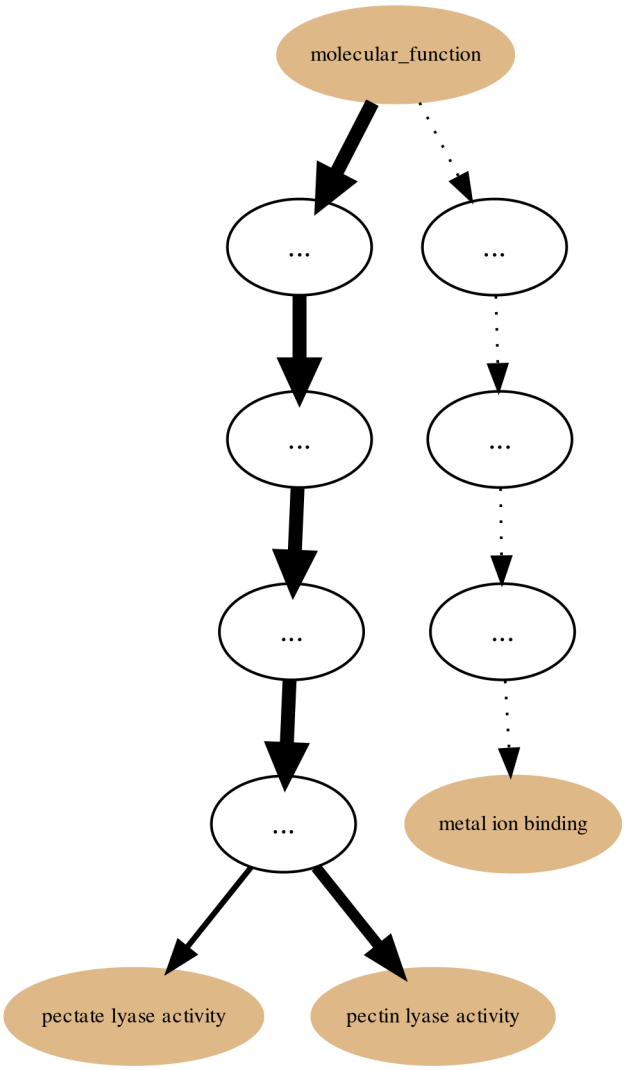


Figure 5.9: *Annotation graph subsuming the PL1 (within the CAZy Collection) Set GO molecular function sub-ontology annotations without electronic annotations (IEA).*

The PL8 family is the third largest PL family on the CAZy database. The *molecular function* sub-ontology annotation graph (using all Evidence Codes) for

the PL8 Set (amounting to 197 proteins) and using the whole CAZy Collection as statistical background was generated. Table 5.2 shows the term annotation occurrence numbers, IC-based scores and p-values for the enriched terms in Set PL8. The top three statistically significant terms are also the most representative in terms of Information Content as can be seen by the values of the IC-based score. Among these three, the term *carbon-oxygen lyase activity, acting on polysaccharides* has the higher score (0.426). When considering that score in conjunction with the "annotation flow" shown on Figure 5.10, it can be seen that about 70% of the proteins are not annotated beyond this term thus making it a good "pivot point" to attempt annotation extension. The remaining proteins that are annotated with terms that are its descendants are mostly (44 proteins) annotated to the term *hyaluronate lyase activity*, the third most IC significant term (IC-based score = 0.147). Furthermore, family PL8 has 3 sub-families (1-3) (Lombard *et al.*, 2010), which in this set is constituted by 88, 33, 4 proteins respectively. Additionally, there are still 77 remaining proteins in Set PL8 that are not classified into any of these aforementioned sub-families. The examination of PL8 sub-family 1 shows that it is characterized by (27) proteins that are annotated to the *hyaluronate lyase activity* term (this term is enriched in the sub-family using the family as background). On the other hand, sub-families 2 and 3 are scarcely annotated beyond the term *carbon-oxygen lyase activity, acting on polysaccharides* and thus do not provide statistical support for what are their most specific representative activities. Hence, further existing annotation would be required for the members of sub-families 2 and 3 in order to assess a more specific functional profile for them.

Even though the web application GRYFUN was designed to be focused for the purpose of protein family coherence measuring and annotation extension assistance, it can also be used for the more common instances where term enrichment is often used, i.e. differential expression gene lists. Therefore, a small-scale microarray study of differential gene expression in human native nasal epithelial cells from five F508del-homozygous cystic fibrosis (CF) patients vs. five control individuals (Clarke *et al.*, 2013) was used for analysis in GRYFUN. For this analysis the genes up-regulated 2-fold or more in CF samples compared to controls were converted into Uniprot accession IDs (n=150), and this single set was then run

term name	occ	IC-based score	p-value
carbon-oxygen lyase activity, acting on polysaccharides	196	0.426	$< 2.531 \times 10^{-225}$
carbohydrate binding	191	0.268	2.531×10^{-225}
hyaluronate lyase activity	44	0.147	2.081×10^{-124}
chondroitin AC lyase activity	3	0.012	3.996×10^{-009}
xanthan lyase activity	2	0.010	2.531×10^{-006}
chondroitin-sulfate-ABC exolyase activity	2	0.009	2.531×10^{-006}
heparin lyase activity	2	0.008	7.042×10^{-005}
chondroitin-sulfate-ABC endolyase activity	1	0.004	1.595×10^{-003}
acharan sulfate lyase activity	1	0.005	1.595×10^{-003}
chondroitin B lyase activity	1	0.005	3.187×10^{-003}
phosphatidylinositol phospholipase C activity	1	0.003	6.365×10^{-003}
metal ion binding	7	0.004	9.978×10^{-003}

Table 5.2: *Term annotated occurrence (occ) number, IC-based term score and enrichment p-values for the PL8 Set significant terms (alpha = 0.01) while using the complete CAZy Collection as background.*

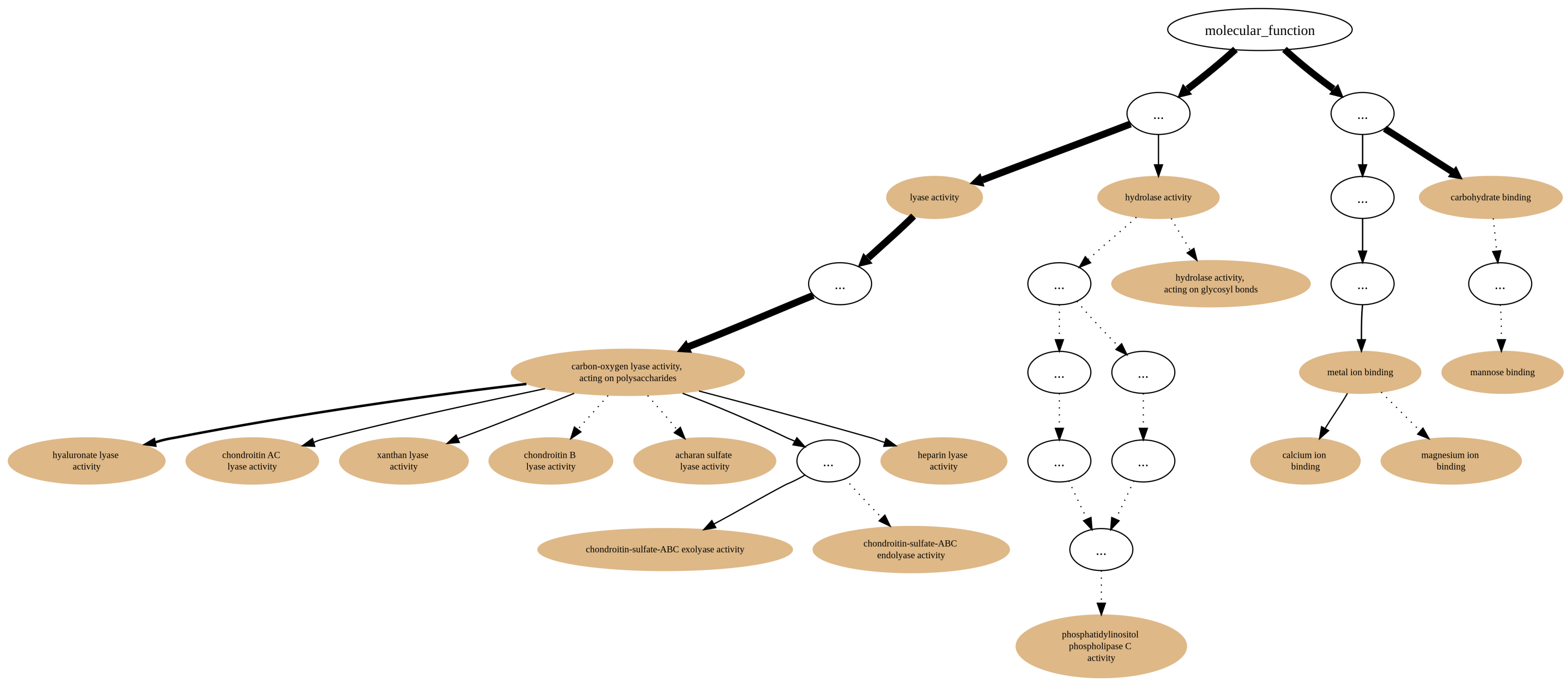


Figure 5.10: *Annotation graph subsuming the PL8 (CAZy family) Set GO molecular function sub-ontology annotations.*

against a background of Uniprot IDs (n=9083) converted from named genes on the Affymetrix HsAirway micro-array used in the study. The results obtained with GRYFUN were then compared with those obtained for the same gene list using two other GO term enrichment software platforms; GOrilla (Eden *et al.*, 2009), using the same background as GRYFUN, and DAVID (Huang *et al.*, 2009), using the default *H. sapiens* background.

GRYFUN identified 90 GO terms (out of 2006) considered statistically significant (alpha = 0.01) annotating the 150 protein identifiers submitted, compared to the 5 identified by DAVID at the same significance level, and the 56 identified by GOrilla (default alpha = 0.001). Some terms were among the most significant identified by all three platforms, while others were only considered significant by one or two platforms, and there were some variations in the number of genes identified as annotated under specific GO terms (see Table 5.3). The variations in the number of annotation occurrences for each term stem from the fact that each of the enrichment tools does not rely exactly on the same releases of annotation databases. GRYFUN also identified as being enriched several GO terms of definite biological significance in the pathophysiology of CF (eg, *Positive Regulation of Cell Differentiation*, *Programmed Cell Death*) which were undetected by other platforms (see Table 5.3).

The relevance of enriched processes to CF, or any other condition being studied, has for GRYFUN as for other enrichment platforms including GOrilla and DAVID, to be assessed by the user, based on knowledge of the processes involved. Additionally, when analysing micro-array data such as this, where there is a high number of biological processes involved, the use of the occurrence number (on statistically enriched terms) can be a quick indicator of how general a process might be. Among the 90 significant GO terms identified by GRYFUN in the used CF data set, approximately 30 had occurrence numbers between 5 and 15, and represented the most functionally relevant, including some of those shown in Table 5.3. The identification of significantly enriched processes by DAVID and not by GRYFUN or GOrilla, and vice versa, may result from the different backgrounds used: the default DAVID background is composed of all human genes with at least one annotation in the category being analysed, whereas the GRYFUN/GOrilla backgrounds are user-defined, in this case being composed of genes

Terms (biological process)	GRYFUN		DAVID		GOrilla	
	p-value	occ	p-value	occ	p-value	occ
Response to Wounding	3.6×10^{-08}	12	3.2×10^{-03}	13	2.1×10^{-04}	11
Immune Response	2.2×10^{-05}	15	4.0×10^{-03}	15	8.8×10^{-05}	21
RNA Biosynthetic Process	$< 1.0 \times 10^{-13}$	15	Not found		Not found	
Programmed Cell Death	$< 1.0 \times 10^{-13}$	11	Not found		Not found	
Positive Regulation of Cell Differentiation	2.7×10^{-06}	14	Not found		3.0×10^{-06}	20
Negative Regulation of Cell Communication	Not found		7.5×10^{-03}	8	9.6×10^{-04}	19
Inflammatory Response	Not found		9.6×10^{-03}	9	3.1×10^{-04}	11
Ectoderm Development	Not found		9.6×10^{-03}	7	Not found	

Table 5.3: *Comparison of GO term enrichment analyses of micro-array data by GRYFUN, DAVID and GOrilla. Selected examples of GO terms found to be enriched in list of differentially expressed genes (upregulated in cystic fibrosis nasal epithelium (Clarke* et al., *2013)) by GRYFUN, DAVID or GOrilla. Occurrence (occ) numbers and p-values are shown. "Not found" means the GO term was not considered significant.*

represented on the micro-array for which a UniProt accession number or Gene Symbol (respectively) was available. Future implementation of a default genome-wide background might standardize the results of enrichment analyses, but the greater number of significant terms produced by GRYFUN in the present analysis could nevertheless prove useful in generating functional hypotheses. For example, the three GO terms identified by GRYFUN in Table 5.3, of which only one was found by GOrilla and none by DAVID (*RNA Biosynthetic Process*, *Programmed Cell Death*, *Positive Regulation of Cell Differentiation*), all have important roles in CF-mediated airway pathology (Booton & Lindsay, 2014; Hajj *et al.*, 2007; Soleti *et al.*, 2013).

The dataset generated here is subsumed by over two thousand (highly interconnected) GO terms (in the biological function sub-ontology), that in turn, renders an extremely complex interactive graph of difficult navigation and interpretation. This is a limitation of the current graph rendering engine used in GRYFUN. However, GRYFUN provides the possibility to download the underlying graph file (by pressing the button "Export .dot" on the Set header) which produces a file that can then be opened with a suitable external viewer application such as the free and cross-platform ZGRViewer (http://zvtm.sourceforge.net/zgrviewer.html). In the future, it is planned to implement additional strategies that help deal with very big graphs, such as pre-rendering additional filters and partial iterative graph loading. Notwithstanding, when a graph of difficult interpretation (due to number of nodes and edges) is generated it is currently possible to immediately perform *re-rooting* operations from the associated term table while guided by the presented statistics. These *re-rooting* operations will then result in smaller and more interpretable branches of the original graph.

Summary

In this chapter the developed web application GRYFUN was presented with its features described and demonstrated with biological examples. This web ap-

plication implements most of this thesis proposed functional coherence analysis module, joining graph visualization and term enrichment analysis and lacking only the local implementation of additional coherence metrics.

The following chapter describes a set of example analyses using the whole current implementation of the whole proposed functional coherence analysis and annotation extension framework.

Chapter 6

Framework Assessement

This thesis proposes a modular framework, presented previously in Chapter 4 and summarized by Figure 4.11, for handling functional annotation coherence analysis of protein families (sets) and extending annotation within those families. Previous validation assays and demonstrations described in Chapter 4 and Chapter 5 have been done to attest to the validity of the individual module implementations proposed in this thesis. In this current chapter, further analytical examples are used to verify the efficiency and validity of the complete pipeline of the specific module instantiations implemented for this thesis.

6.1 MEROPS

The MEROPS database has hierarchical classifications in which homologous sets of peptidases and protein inhibitors are grouped into protein species, which themselves are grouped into families, which are in turn grouped into clans (Rawlings *et al.*, 2012). Thus, the classification in this database makes it convenient for functional annotation analysis using GRYFUN. All the UniProt identifiers (n=93124) from this database (MEROPS release 9.9) were extracted. MEROPS families with four or more UniProt identifiers were recreated as Sets in a GRYFUN Collection that ended up being composed of 238 Sets. A few family sets where chosen based on their annotation graphs and available information available in order to present an extended analysis on how to representatively use the framework. The analysis performed on those sets is described below.

Family set: A2

Set (family) A2 was chosen from the MEROPS Collection and its annotation graph and associated statistics for the GO *molecular function* sub-ontology were generated. The annotation graph is shown in Figure 6.1 while Table 6.1 displays the associated statistics. The MEROPS website (http://merops.sanger.ac.uk/cgi-bin/famsum?family=A2) describes peptidase family A2 as containing "endopeptidases with catalytic sites of aspartic type". Table 6.1 shows that the enriched term *aspartic-type endopeptidase activity* is the most specific and prevalent one (IC-based term score = 0.222; annotates 77% of the Set), thus supporting the MEROPS family classification for this Set. Additional high-scoring terms in this Set are *RNA-directed DNA polymerase activity*, *RNA-DNA hybrid ribonuclease activity* and *RNA binding* all of which are functions inherently related to the reported family type, HIV-1 retropepsin as can be verified by the available domain knowledge.

term name	score	p-value
RNA-directed DNA polymerase activity	0.194	7.24×10^{-271}
RNA-DNA hybrid ribonuclease activity	0.158	5.01×10^{-171}
RNA binding	0.106	3.27×10^{-165}
aspartic-type endopeptidase activity	0.222	1.05×10^{-131}
exoribonuclease H activity	0.077	4.80×10^{-045}
nucleic acid binding	0.074	1.82×10^{-041}
DNA binding	0.035	4.86×10^{-035}
DNA-directed DNA polymerase activity	0.034	5.18×10^{-032}
zinc ion binding	0.091	3.36×10^{-031}
structural molecule activity	0.039	3.86×10^{-022}
structural constituent of virion	0.013	1.70×10^{-014}
dUTP diphosphatase activity	0.012	4.88×10^{-011}
phosphotransferase activity, alcohol group as acceptor	0.001	3.57×10^{-003}
transferase activity, transferring hexosyl groups	0.002	3.57×10^{-003}

Table 6.1: *Annotation IC-based term score and enrichment p-values for the A2 Set significant terms (alpha = 0.01) while using the complete collection of MEROPS Collection as background.*

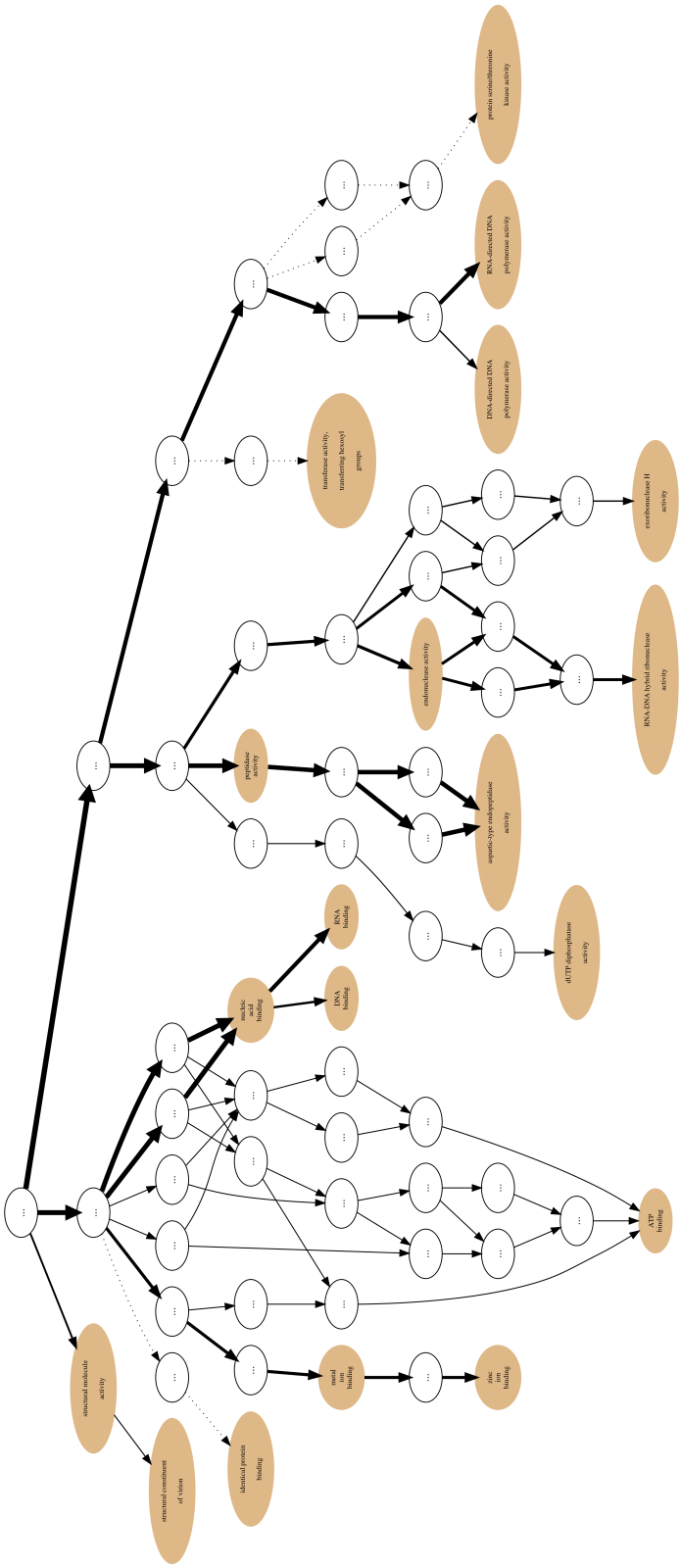


Figure 6.1: *Annotation graph subsuming the A2 Set (MEROPS Collection) GO molecular function sub-ontology annotations.*

Furthermore, the annotation graph for Set A2 is easy to navigate and there are several "annotation flow" paths flowing towards specific relevant terms. Therefore, according to the proposed framework, there are still potential annotation extension opportunities down each of these paths since none of the terms found significant annotates all the proteins in the A2 Set. With that in mind, the top five terms from Table 6.1 were chosen to create HMM profiles to be assayed as previously described in Section 4.4. The results for this assay are shown in Table 6.2.

term name	annotations	recall	σ
RNA-directed DNA polymerase activity	109	0.816	0.061
RNA-DNA hybrid ribonuclease activity	68	0.773	0.086
RNA binding	105	0.789	0.084
aspartic-type endopeptidase activity	120	0.629	0.121
exoribonuclease H activity	16	0.520	0.160

Table 6.2: *Recall and respective standard deviations for the repeated random subsampling validation procedure on a set of chosen (top five) significant terms from Set A2 from the MEROPS99 Collection.*

The obtained recall results, considering that no MSA/HMM manual tuning was done by a domain expert or otherwise, are reasonably high. Curiously enough the recall ranking nearly matches the enrichment ranking for Set A2. It can be inferred that the automatically generated MSA/HMM profile best captures the domain responsible for the *RNA-directed DNA polymerase activity*. Although a direct correlation between MSA/HMM profile size and recall cannot be observed, the MSA/HMM profile generated from higher numbers of sequences has decreased the standard deviations when compared to previous results (Table 4.5). It is then not surprising to find the HMM profile for term *exoribonuclease H activity* having the lowest recall (0.520) considering that it derives from a MSA generated out of only 11 sequences. On the other hand, the second lowest recall value (0.629) comes from assaying of the HMM profile for term *aspartic-type endopeptidase activity* despite it being generated from the largest MSA (80 sequences) in this Set A2 assay. According to the A2 family classification in the MEROPS database

this activity should be expected to annotate all of its proteins. However, the relatively poor recall suggests that the currently implemented annotation extension module would benefit from additional intervention by a specialized curator for verifying and adjusting the created MSA and subsequent HMM profiles. The focus of this thesis does not fall on the optimization of the annotation procedure. Nevertheless, it will be interesting for future developments of this module to team up with a MSA/HMM specialist in order to create features, such as alignment visualization plugins enabling the editing of the automatically generated MSAs to iteratively access and adjust the HMMER program parameters. This will be potentiating the discovery and segregation of individual domains or functional modules directly associated with individual annotation terms.

Additionally, the whole A2 Set and separately a subset of its proteins annotated to the term *peptidase activity* was submitted to the metrics assayed in Chapter 4 and the results presented in Table 6.3.

metrics	A2 (full) 162 proteins	A2 (subset) 126 proteins
Agreement	0.313	0.378
simUI	0.499	0.566
simGIC	0.414	0.507
mUI	0.380	0.436
mGIC	0.380	0.447
GS^2	0.835	0.875
Completeness (leaf-assumption)	98.77%	99.21%
Completeness (IC threshold assumption)	9.88%	12.70%

Table 6.3: *Agreement, coherence and completeness metrics for Set A2 and its subset of GO term peptidase activity annotated proteins.*

As expected, the selected subset reports higher values for all the tested metrics when compared with the results for the complete A2 Set. As evidenced by the graph in Figure 6.1 and supported by the enrichment results in Table 6.1 the A2 Set is fairly heterogeneously annotated with several terms found significant. Accordingly to the leaf-assumption completeness metric results, the annotation of this set should be complete, that is nearly every protein is at least annotated

to one leaf-term. However, given the multi-functional nature of proteins that is not actually expected to hold to the absolute extent, and as previously mentioned this is a naive metric. Nevertheless, the reported results for the GS^2 are quite elevated considering the distribution of annotations in the Set. Again, as previously reported, GS^2 has a small resolution biased for higher scores. On the other hand, the remaining scores are much closer together and seem to represent the actual annotation state for the A2 Set.

Family set: C15

The C15 family from MEROPS containing peptidases of cysteine catalytic type was also selected for analysis with the proposed framework. According to information provided by MEROPS (http://merops.sanger.ac.uk/cgi-bin/famsum?family=C15) its only known activity is the removal of a pyroglutamate (pGlu) residue from the N-terminus of a peptide. The sequences from this family imported into the MEROPS Collection on GRYFUN resulted in the annotation graph in Figure 6.2. The information regarding this family that is provided in the MEROPS database is easily confirmed by the generated graph where the annotation flow can be seen basically flowing towards just two leaf nodes, a) *cysteine-type peptidade activity* and b) *pyroglutamyl-peptidase activity*. This is further reinforced by the enrichment results shown in Table 6.4 where only these two leaf terms are found to be enriched. Most (251 out of 265) of the proteins in this Set are annotated with term *cysteine-type peptidade activity* which is characteristic of this class of peptidase whilst term *pyroglutamyl-peptidase activity* annotates over 60% of the proteins in C15.

term name	score	occ	p-value
pyroglutamyl-peptidase activity	0.533	164	$<5.70\times10^{-248}$
cysteine-type peptidase activity	0.345	251	5.70×10^{-248}

Table 6.4: *Annotation IC-based term score, numbers of annotations (occ) and enrichment p-values for the C15 Set significant terms (alpha = 0.01) while using the complete MEROPS Collection as background.*

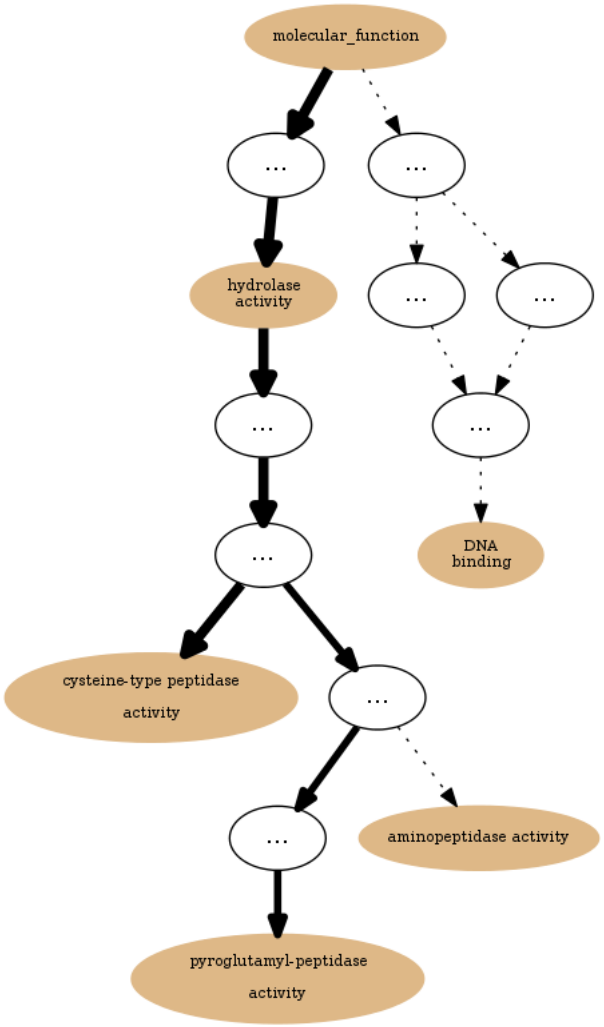


Figure 6.2: *Annotation graph subsuming the C15 Set (MEROPS Collection) GO molecular function sub-ontology annotations.*

The agreement and coherence metrics were also run for this Set and the results presented in Table 6.5.

	Agreement	simUI	simGIC	mUI	mGIC	GS^2
C15	0.515	0.800	0.661	0.773	0.724	0.960

Table 6.5: *Agreement and coherence metrics for MEROPS Set C15.*

All the results are within the expected ranges considering the previous assays

and they reflect the particular properties of each metric. Notably, the Agreement in this Set is only around 50% because nearly half the proteins are not annotated with the term *pyroglutamyl-peptidase activity* which is, as reported, the only known specific activity for the C15 family. It is then important to strive to extend the annotation of this term within this family.

Until now only the recall metric has been used for measuring the retrieving ability of relevant (annotated with the extension term candidate) protein sequences by the implementation of the annotation extension module used in this thesis. That is because an open world assumption was being considered due to the fact that a protein can have any number of annotations. Additionally, considering the fact of dealing with protein families (or sets) that are incompletely annotated the possibility that completeness can only be achieved through total annotation agreement cannot be discarded. However, it is important to ascertain the discriminatory power of the implemented module, thus the procedures previously run to obtain the recalls results were modified to obtain precision results too. As previously mentioned any protein can have any number of non-exclusive annotation terms. However, now a closed world assumption is made where random proteins outside any given protein family and not annotated with a particular term will not possess the particular function that the term describes. Therefore, the previously used methodology for repeated random sub-sampling validation was altered to randomly replace half of the test set by random proteins not annotated by the candidate extension term. This modified methodology was run on the HMM profiles created for the *cysteine-type peptidase activity* and *pyroglutamyl-peptidase activity* annotation candidate terms in Set C15 and the results presented in Table 6.6.

	cysteine-type peptidase activity	pyroglutamyl-peptidase activity
precision	1.000	1.000
recall	0.743	0.833
F-score	0.852	0.909

Table 6.6: *Precision, recall and F-score for the two candidate annotation extension terms in the MEROPS Set C15.*

Both of the annotation extension candidate terms HMM profiles achieve a precision of 100%, that is, only protein sequences actually annotated with any of the terms were identified as such. This is crucial for high quality annotation because although it is desirable to increase annotation coverage that should not be done at the cost of precision loss. The recalls are also relatively high when compared with previously shown results for assayed families.

Family set: N6

The N6 family is of the asparagine-type peptidase activity and is described as including autoprocessing endopeptidases. However, despite being constituted by a set of 182 proteins it is very scarcely annotated, only 3 of those proteins have GO term annotations of the molecular function ontology. Figure 6.3 displays the annotation graph where the scarcity of annotation is evident. In cases like these the proposed framework is able to provide little assistance and other methodologies must be employed to increment the annotation coverage of the family under analysis.

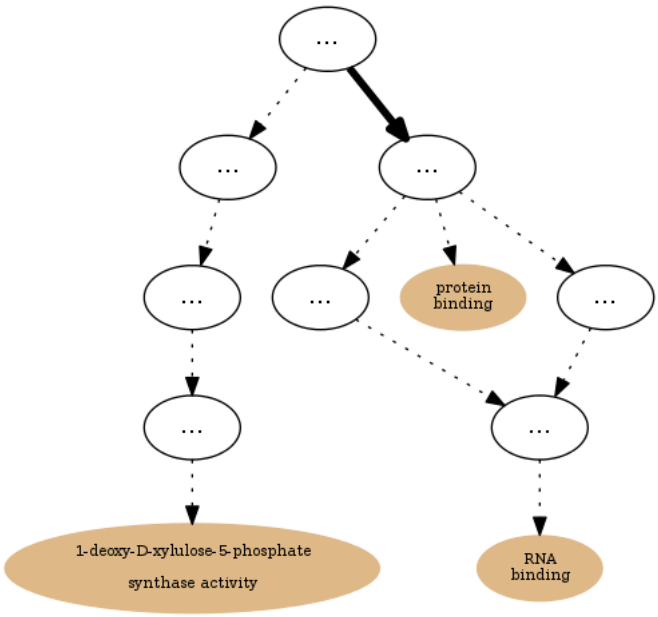


Figure 6.3: *Annotation graph subsuming the N6 Set (MEROPS Collection) GO molecular function sub-ontology annotations.*

6.2 CAZy

The PL class CAZy families were used to develop the methodology assayed in this thesis, because despite their sub-optimal annotation state they are the best studied group of CAZy families and support was offered by the CAZy curator team. However, there are three other enzymatic classes (GH, GT and CE) in the CAZy database. Therefore, the resulting analysis of an example family from each of those classes is presented below.

Family set: GH70

According to cazypedia, a resource maintained by expert curators on carbohydrate-active enzymes, the CAZy family GH70 (http://www.cazypedia.org/index.php/Glycoside_Hydrolase_Family_70) is composed of transglucosylases, also known as glucosyltransferases or glucansucrases. The annotation graph for GH70 molecular function ontology was then generated and displayed in Figure 6.4 while its enrichment results are presented in Table 6.7 and the agreement and results for the coherence metrics in Table 6.8.

term name	annotations	p-value
glucosyltransferase activity	102	1.04×10^{-138}
dextransucrase activity	37	6.05×10^{-118}
oligosaccharide 4-alpha-D-glucosyltransferase activity	2	5.91×10^{-007}
1,4-alpha-glucan 6-alpha-glucosyltransferase activity	2	5.91×10^{-007}
alternansucrase activity	1	7.73×10^{-004}

Table 6.7: *Number of annotations and enrichment p-values for the GH70 Set significant terms (alpha = 0.01) while using the complete CAZy Collection as background.*

All the terms marked as enriched match the information present in cazypedia, thus confirming that in this case statistical relevance matches biological relevance. However, three of those terms (bottom three at Table 6.7 or Figure 6.4) are annotated to either one or two proteins. Consequently, these sparse, yet significant, annotations lower the value of the mUI and mGIC metrics, when compared with the values from the metrics they derive from (simUI and simGIC, respectively).

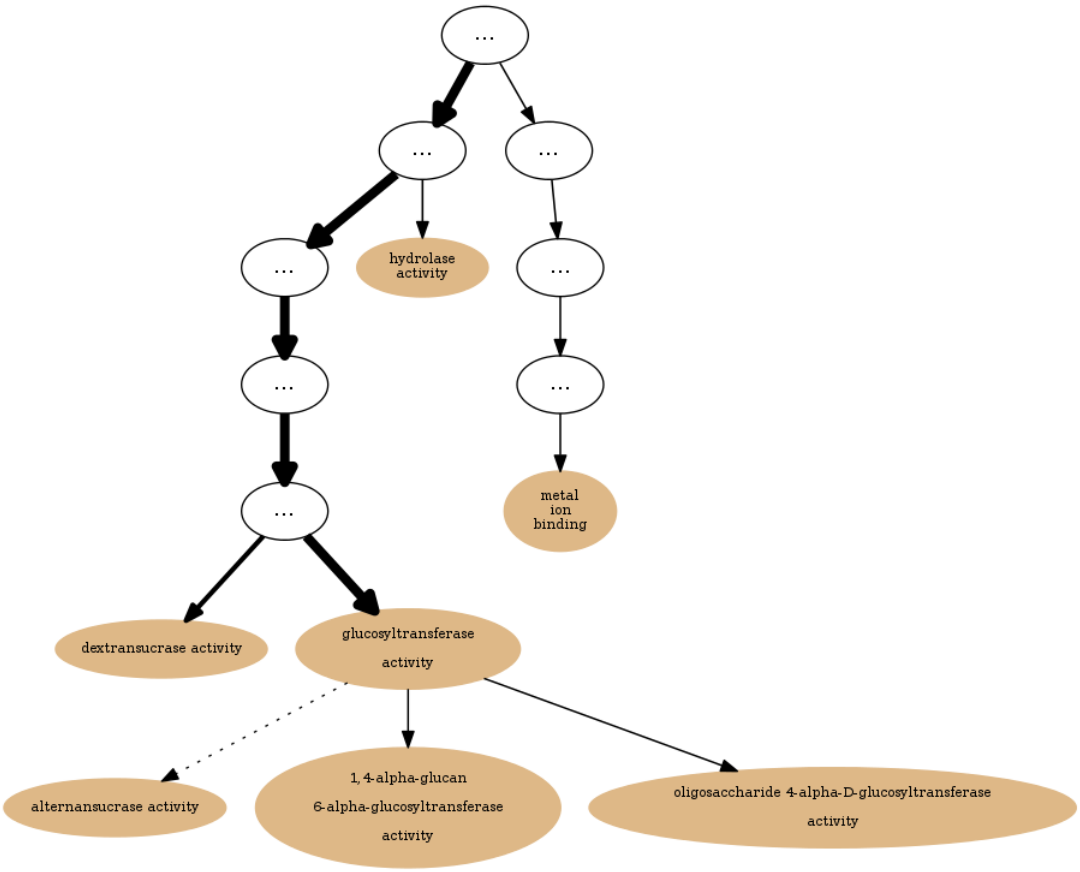


Figure 6.4: *Annotation graph subsuming the GH70 Set (CAZy Collection) GO molecular function sub-ontology annotations.*

	Agreement	simUI	simGIC	mUI	mGIC	GS^2
GH70	0.442	0.888	0.770	0.714	0.656	0.984

Table 6.8: *Agreement and coherence metrics for CAZy Set GH70.*

Additionally, the low number of proteins annotated with these three terms in Set GH70 does not allow to apply the proposed annotation extension methods on them. However, their parent term *glucosyltransferase activity* annotates all the proteins in the Set as would be expected from the family description. Since currently there are no proteins without this annotation it would not make sense to try annotation extension of this term unless when considering to add new proteins sequences to the GH70 family. On the other hand, the term *dextransucrase*

activity is also reported to be characteristic of the GH70 family while annotating less than half of its proteins. Hence, HMM profiles were generated for both these terms and their precision and recall ascertained via the previously described repeated random sub-sampling validation method and the results presented in Table 6.9.

term name	precision	recall	σ	F-score
glucosyltransferase activity	1.000	0.827	0.085	0.905
dextransucrase activity	1.000	0.700	0.180	0.824

Table 6.9: *Precision, recall and F-score for the two candidate annotation extension terms in the CAZy Set GH70.*

For these HMM profiles the precision is optimal (100%) and the recall is high in the order of previously obtained results. However, for the *dextransucrase activity* the HMM profile average recall is 70% with a standard deviation of 18%. This again results from using a smaller number of sequences to generate an MSA and consequently also a small testing set that is further reduced by the replacement of half its proteins with random ones as previously described.

Family set: GT44

As before, the molecular function ontology annotation graph (Figure 6.5) for the Set GT44 was generated along with its term enrichment (Table 6.10).

The annotation flow observed in Figure 6.5 immediately indicates which are the likely characteristic activities in this Set. Most of the annotation flow goes to terms *cysteine-type endopeptidase activity* and *transferase activity, transferring glycosyl groups*. These terms are confirmed to be enriched by the results presented in Table 6.10 with an additional term, *dioxygenase activity*. However, this last term only annotates one protein and so, as described before, there is no possible follow through with the proposed framework. Nevertheless, the former terms can be used as annotation extension term candidates and despite having low IC score and thus not being very informative. Taking in consideration that GT44 belongs to the GT class it would have been optimal to find more informative children of the term *transferase activity, transferring glycosyl groups* annotating

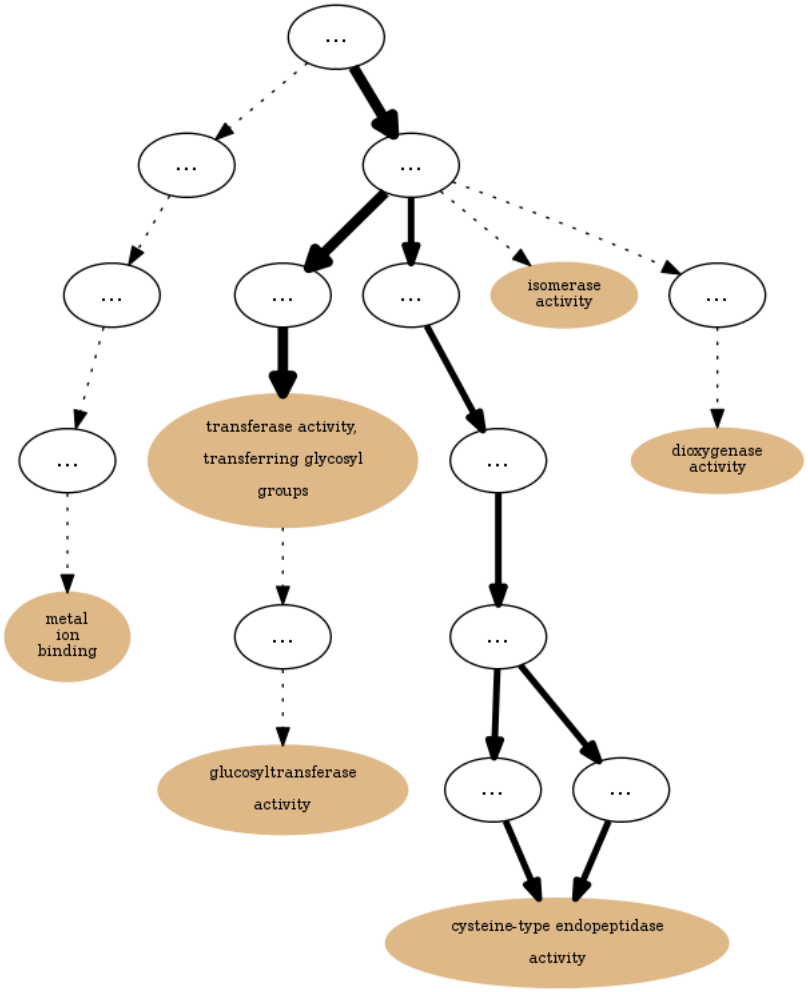


Figure 6.5: *Annotation graph subsuming the GT44 Set (CAZy Collection) GO molecular function sub-ontology annotations.*

this family (other than the single *glucosyltransferase activity* annotation). The HMM profiles were created and assayed as previously described, and despite the results demonstrating 100% precision for the generated HMM models their recall was relatively low, with 0.76±0.20 for *cysteine-type endopeptidase activity* and 0.64±0.17 for *transferase activity, transferring glycosyl groups*. Again, the lower number of sequences available to generate the HMM profiles and testing them can explain these relatively lower recall values with reasonably high standard deviations.

term name	IC score	occ	p-value
cysteine-type endopeptidase activity	0.381	30	5.14×10^{-94}
transferase activity, transferring glycosyl groups	0.235	53	1.22×10^{-30}
dioxygenase activity	0.314	1	3.69×10^{-03}

Table 6.10: *IC score, number of annotations (occ) and enrichment p-values for the GT44 Set significant terms (alpha = 0.01) while using the complete CAZy Collection as background.*

Unlike Set GH70 for Set GT44 the metrics mUI and mGIC score higher (0.703 and 0.652 respectively) than the metrics they derived from (0.666 and 0.567 respectively) because of the evident higher annotation homogeneity among the significant terms in this Set. Considering that for both these aforementioned Sets there are annotations that despite their biological relevance do not allow for extrapolation a mechanism can be implemented to excise these annotations from the coherence accounting and shift them to the incompleteness accounting. Thus, the implementation of a future plugin for the iterative selection/pruning of terms can be an useful for an expert curator to discard terms (even if statistically enriched) either deemed insufficient or unrelated via domain knowledge.

Family set: CE7

The last example is the CE7 Set, a family belonging to the Carbohydrate Esterases (CE) class in the CAZy database. Its annotation graph (Figure 6.6) shows that the annotations flow mostly toward the term *cephalosporin-C deacetylase activity*. The enrichment results (Table 6.11) confirm this term as enriched along with two other, which as in the previous example can not be used for annotation extension due to them annotating only a small amount of proteins. Compared to previous assays even the number of annotations for the term *cephalosporin-C deacetylase activity* may be low (19) but an attempt at measuring the extension capability of its derived HMM profile was performed. As with the previous profiles, a precision of 100% was achieved with the essay, however the recall only scored 0.367±0.233. It should be noted that only 13 sequences were used to con-

struct the MSA/HMM profile while 6 for testing, with 3 of them being randomly replaced. Hence a new assay was performed, this time without the random replacement and thus having 6 relevant protein sequences for testing. That resulted in an increase for the obtained recall score to 0.450±0.269.

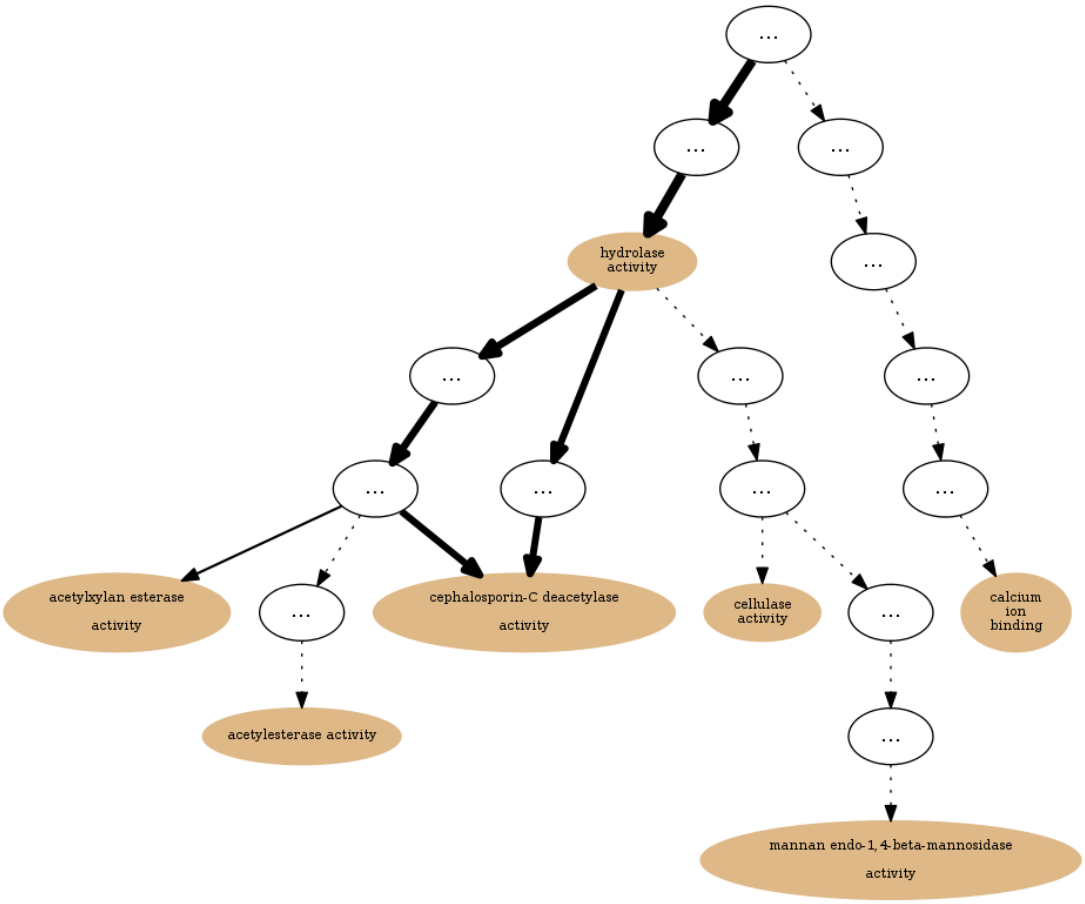


Figure 6.6: *Annotation graph subsuming the CE7 Set (CAZy Collection) GO molecular function sub-ontology annotations.*

The results obtained so far point to the need for having a reasonable number of protein sequences to generate MSA/HMM profiles that are able to recall acceptable percentages of sequences given that is the purpose of the annotation extension module. However, it should be noted that all of the HMM profiles generated and assayed so far were able to achieve 100% precision, which is crucial

because it prevents error propagation.

term name	IC score	occ	p-value
cephalosporin-C deacetylase activity	0.731	19	2.68×10^{-74}
acetylxylan esterase activity	0.766	3	1.40×10^{-08}
acetylesterase activity	0.688	1	2.00×10^{-03}

Table 6.11: *IC score, number of annotations (occ) and enrichment p-values for the CE7 Set significant terms (alpha = 0.01) while using the complete CAZy Collection as background.*

Summary

This chapter has demonstrated through examples the feasibility of the proposed framework. It reciprocally helps domain experts by automating some procedures and benefits from the intervention of the same experts to fine tune the modules. The framework is modular and designed so that other methodologies can be inserted or replace the ones implemented in this thesis. The task of analysing the functional annotation coherence of any given set can be a complex task where a single metric is too reductive not taking into account all the dimensions involved in protein annotation. Therefore, here was suggested and assayed beyond novel coherence metrics, a complete integrated framework making use of semantic similarity metrics, graph visualization, term enrichment techniques and sequence alignment methods.

Chapter 7

Conclusions

The discrepancy between the output pace of high-throughput sequencing methods and the functional annotation efforts has led to an heterogeneous annotation landscape both in terms of coverage reach and the functional specificity of that coverage. Ideally, proteins should be annotated in a way that fully describes their functional activities. However, even within the boundaries of current knowledge, this is seldom the case. Thus, when trying to assert the functional coherence of proteins, such as families, based on their functional annotations this heterogeneity of functional annotation becomes a greater issue. Annotation incompleteness can lead to false interpretations about the existing functional inter-similarity within any given protein set (or family). In order to avoid erroneous interpretations on heterogeneous protein sets or families (in terms of annotation specificity), functional comparisons are usually done at conservative levels. Therefore, there is a dual need to address the issues of extending the annotation coverage in reach and specificity for the protein universe, but also there is the need to accurately measure functional coherence in groups of proteins despite the current heterogeneous annotation landscape.

The work presented in this thesis proposes a modular framework that employs techniques to handle both measuring of functional annotation coherence in protein families (or other functionally related sets) and a methodology to propose annotation extension for proteins of those same families. Instantiations of modules needed to build this framework were implemented, tested and validated.

In the Introduction chapter (Chapter 1) the **Hypothesis** was introduced as: *It is possible to extend functional annotations in protein families with the assistance of adequate functional coherence analysis considering that families are expertly collected knowledgebases.*

This thesis assessed this hypothesis through the implementation of the proposed modular framework that was developed and tested, asserting and measuring protein family functional annotation coherence resorting to semantic similarity metrics and enrichment techniques with further aid of graph visualization.

7.1 Functional coherence metrics

As presented in Chapter 4 a module for assessing functional coherence in protein families was developed. This module relies on the hybrid use of term enrichment techniques and semantic similarity metrics. Two novel hybrid metrics mUI and mGIC were developed to capture functional representativeness and assess local coherence within protein sets based on subsets of functions deemed representative. Other metrics were assayed alongside these novel metrics and a methodology of functional coherence analysis in protein families (sets) was demonstrated using the proposed module. This was achieved by integrating term enrichment analysis, graph visualization and semantic similarity-based metrics.

7.2 GRYFUN

The web application GRYFUN, present in Chapter 5 was developed for the visualization of GO annotation graphs of protein sets. It was designed for aiding in annotation coherence and cohesiveness analysis and annotation extension assessments within under-annotated protein sets. Additionally, GRYFUN was designed to be extensible in order to accommodate all the modules necessary for the proposed framework. Currently, it accepts lists of UniProt accession numbers in order to create user-defined protein sets enabling subsequent annotation graph visualization along with term enrichment and support statistics. GRYFUN is freely and publicly available at `http://xldb.di.fc.ul.pt/gryfun/` and

also all of its code available (under an MIT license) on a public GIT server (https://bitbucket.org/hpbastos/gryfunserver/) so that anyone can modify, contribute to or simply deploy GRYFUN in their own servers.

7.3 Annotation extension in protein families

The module for extension in protein families connects with the module for the functional coherence metrics and takes as input proteins annotated with enriched terms in protein families (sets). In this work, this module was implemented as a module wrapper library that conducts automated multiple alignment of the previously selected protein family subsets and builds Hidden Markov Model profiles that can then be used as classifiers to extend a given annotation into other protein sequences in a family, or even external sequences. The assaying of this module demonstrated its high precision (100%) for all assayed models and also reasonable to high recall values when a sufficient number of protein sequences were used to generate the HMM profiles.

7.4 Limitations and future work

The major contribution of this work is the proposed annotation extension framework that provides a methodology and set of tools that allow to better harness knowledge from within specialized protein databases. Therefore, as demonstrated, the developed methodology enables the extraction of knowledge from protein families (sets) within collections which in turn enables annotation extension within those same families when certain conditions of sub-annotation are met as previously described. Despite the demonstrated feasibility of the proposed framework it is not without limitations. It is shown in this work that the use of a single metric for asserting the functional coherence within a protein set is reductive, because of the multi-dimensional incompleteness of the annotation space. The developed web application GRYFUN already integrates the graph visualization and term enrichment component of the annotation coherence module. In the future, the novel semantic-based and enrichment-based mUI and mGIC hybrid metrics will also be integrated into GRYFUN thus completely encapsulating the

first module of the framework into a single tool. Additionally, in order to enhance GRYFUN capabilities some features and improvements are either planned or already being implemented. Regarding the input options it is planned to extend the number of identifiers allowed and handled by GRYFUN and even the possibility of uploading customised annotation mapping files. Furthermore, features that enable the direct handling of data, such as directly creating new sets out of subsets selected from already existing sets and the ability of selecting (immutable) preloaded Collections as statistical backgrounds is also planned. Concerning graph output the plan is to implement rendering options and filtering options in order to enhance graph readability and interpretation, which is especially useful to deal with larger annotation graphs.

The annotation extension module of the proposed framework has the most margin for improvement. It is currently implemented as a binding library and a series of scripts. In the future, those scripts can be developed into a standalone tool able to seamlessly integrate with GRYFUN, as it was originally intended. Among the planned improvements is the development of a MSA editor, that would allow a potential expert curator to make adjustments to the automatically produced MSA. The development of this particular improvement would benefit greatly from the collaboration of expert curators. In addition, just like for GRYFUN all code will be made available so that anyone can modify it or contribute to its development.

Besides the current limitations this work already enables the extension of functional annotation of the protein universe, through the functional annotation coherence analysis of partially (to fully) annotated protein (enzyme) families while using the proposed methodology and developed tools.

Appendix A

Similarity results for randomization assays

Each tested CAZy PL families (PL1-PL12, PL16, PL17 and PL22) were altered by progressively replacing discrete amounts (with 10% increments) of original family proteins with the same amounts of proteins randomly selected from the CAZy database. Each of these created sets was measured with the Agreement, simUI, simGIC, mUI, mGIC and GS^2 metrics.

A. SIMILARITY RESULTS FOR RANDOMIZATION ASSAYS

PL1	Agreement	simUI	simGIC	mUI	mGIC	GS^2
similarity	0.159	0.662	0.498	0.466	0.472	0.914
random proteins	10%		20%		30%	
	similarity	σ	similarity	σ	similarity	σ
Agreement	0.073	0.009	0.056	0.006	0.049	0.005
simUI	0.591	0.005	0.528	0.006	0.476	0.007
simGIC	0.413	0.005	0.337	0.006	0.274	0.006
mUI	0.463	0.008	0.462	0.011	0.457	0.017
mGIC	0.469	0.008	0.468	0.010	0.462	0.017
GS^2	0.882	0.003	0.853	0.004	0.829	0.004
random proteins	40%		50%		60%	
	similarity	σ	similarity	σ	similarity	σ
Agreement	0.044	0.003	0.042	0.003	0.040	0.003
simUI	0.434	0.008	0.400	0.008	0.380	0.007
simGIC	0.223	0.007	0.184	0.006	0.159	0.005
mUI	0.454	0.022	0.453	0.032	0.451	0.039
mGIC	0.460	0.022	0.458	0.031	0.455	0.036
GS^2	0.810	0.005	0.795	0.005	0.785	0.005
random proteins	70%		80%		90%	
	similarity	σ	similarity	σ	similarity	σ
Agreement	0.039	0.003	0.038	0.003	0.038	0.002
simUI	0.367	0.007	0.363	0.007	0.372	0.008
simGIC	0.144	0.005	0.142	0.006	0.152	0.007
mUI	0.461	0.052	0.464	0.066	0.466	0.084
mGIC	0.462	0.050	0.464	0.065	0.465	0.085
GS^2	0.779	0.005	0.777	0.005	0.781	0.006
random proteins	100%					
	similarity	σ				
Agreement	0.037	0.002				
simUI	0.389	0.010				
simGIC	0.175	0.009				
mUI	0.327	0.351				
mGIC	0.325	0.349				
GS^2	0.788	0.006				

Table A.1: *Average similarity results (as measured with the Agreement, simUI, simGIC, mUI, mGIC and GS^2 metrics) and respective standard deviations (σ) for the PL1 family and derived sets with progressive (10% increments; 100 iterations) amounts of original PL1 proteins being replaced by random proteins taken from other CAZy families.*

PL2	Agreement	simUI	simGIC	mUI	mGIC	GS^2
similarity	0.519	0.844	0.676	0.323	0.323	0.979

random proteins	10%		20%		30%	
	similarity	σ	similarity	σ	similarity	σ
Agreement	0.312	0.065	0.240	0.044	0.210	0.038
simUI	0.741	0.012	0.646	0.015	0.545	0.016
simGIC	0.566	0.008	0.468	0.013	0.356	0.010
mUI	0.321	0.010	0.321	0.015	0.319	0.020
mGIC	0.321	0.010	0.321	0.015	0.319	0.020
GS^2	0.929	0.009	0.883	0.012	0.837	0.013

random proteins	40%		50%		60%	
	similarity	σ	similarity	σ	similarity	σ
Agreement	0.187	0.034	0.163	0.026	0.151	0.029
simUI	0.480	0.016	0.414	0.017	0.379	0.020
simGIC	0.287	0.012	0.215	0.013	0.175	0.015
mUI	0.312	0.028	0.309	0.039	0.291	0.051
mGIC	0.312	0.028	0.309	0.039	0.291	0.051
GS^2	0.807	0.015	0.778	0.013	0.765	0.015

random proteins	70%		80%		90%	
	similarity	σ	similarity	σ	similarity	σ
Agreement	0.142	0.021	0.135	0.021	0.128	0.019
simUI	0.357	0.020	0.347	0.026	0.356	0.029
simGIC	0.151	0.015	0.137	0.022	0.146	0.024
mUI	0.290	0.065	0.323	0.152	0.427	0.277
mGIC	0.290	0.065	0.323	0.152	0.426	0.277
GS^2	0.757	0.017	0.755	0.020	0.764	0.021

random proteins	100%	
	similarity	σ
Agreement	0.133	0.024
simUI	0.390	0.028
simGIC	0.176	0.029
mUI	0.213	0.359
mGIC	0.210	0.357
GS^2	0.790	0.017

Table A.2: *Average similarity results (as measured with the Agreement, simUI, simGIC, mUI, mGIC and GS^2 metrics) and respective standard deviations (σ) for the PL2 family and derived sets with progressive (10% increments; 100 iterations) amounts of original PL2 proteins being replaced by random proteins taken from other CAZy families.*

A. SIMILARITY RESULTS FOR RANDOMIZATION ASSAYS

PL3	Agreement	simUI	simGIC	mUI	mGIC	GS^2
similarity	0.308	0.974	0.978	0.991	0.990	0.992
random proteins	10%		20%		30%	
	similarity	σ	similarity	σ	similarity	σ
Agreement	0.134	0.022	0.098	0.017	0.078	0.009
simUI	0.832	0.004	0.708	0.005	0.602	0.006
simGIC	0.798	0.003	0.640	0.004	0.503	0.004
mUI	0.988	0.017	0.990	0.008	0.987	0.015
mGIC	0.987	0.017	0.989	0.008	0.986	0.015
GS^2	0.931	0.004	0.880	0.004	0.837	0.005
random proteins	40%		50%		60%	
	similarity	σ	similarity	σ	similarity	σ
Agreement	0.069	0.008	0.063	0.006	0.058	0.005
simUI	0.515	0.006	0.444	0.006	0.396	0.007
simGIC	0.389	0.004	0.294	0.004	0.226	0.006
mUI	0.984	0.016	0.983	0.023	0.970	0.034
mGIC	0.983	0.017	0.983	0.023	0.969	0.035
GS^2	0.803	0.006	0.777	0.005	0.762	0.006
random proteins	70%		80%		90%	
	similarity	σ	similarity	σ	similarity	σ
Agreement	0.054	0.005	0.051	0.005	0.049	0.004
simUI	0.366	0.009	0.355	0.009	0.362	0.010
simGIC	0.179	0.007	0.154	0.008	0.153	0.009
mUI	0.953	0.052	0.927	0.072	0.792	0.172
mGIC	0.952	0.053	0.927	0.072	0.792	0.172
GS^2	0.754	0.007	0.757	0.007	0.768	0.007
random proteins	100%					
	similarity	σ				
Agreement	0.048	0.004				
simUI	0.388	0.012				
simGIC	0.173	0.011				
mUI	0.320	0.362				
mGIC	0.318	0.361				
GS^2	0.788	0.008				

Table A.3: *Average similarity results (as measured with the Agreement, simUI, simGIC, mUI, mGIC and GS^2 metrics) and respective standard deviations (σ) for the PL3 family and derived sets with progressive (10% increments; 100 iterations) amounts of original PL3 proteins being replaced by random proteins taken from other CAZy families.*

PL4 similarity	Agreement 0.486	simUI 0.722	simGIC 0.600	mUI 0.620	mGIC 0.610	GS^2 0.930
random proteins	10% similarity	σ	20% similarity	σ	30% similarity	σ
Agreement	0.263	0.045	0.218	0.041	0.190	0.040
simUI	0.631	0.009	0.554	0.012	0.488	0.015
simGIC	0.495	0.009	0.402	0.009	0.324	0.010
mUI	0.614	0.021	0.609	0.023	0.603	0.030
mGIC	0.604	0.020	0.599	0.022	0.592	0.031
GS^2	0.884	0.008	0.845	0.010	0.812	0.012
random proteins	40% similarity	σ	50% similarity	σ	60% similarity	σ
Agreement	0.163	0.029	0.151	0.029	0.143	0.025
simUI	0.433	0.014	0.394	0.019	0.366	0.019
simGIC	0.260	0.012	0.210	0.013	0.174	0.014
mUI	0.598	0.045	0.585	0.051	0.583	0.076
mGIC	0.588	0.043	0.574	0.046	0.571	0.072
GS^2	0.787	0.010	0.770	0.014	0.758	0.015
random proteins	70% similarity	σ	80% similarity	σ	90% similarity	σ
Agreement	0.133	0.023	0.123	0.019	0.118	0.016
simUI	0.351	0.021	0.347	0.020	0.361	0.026
simGIC	0.153	0.016	0.142	0.016	0.150	0.022
mUI	0.577	0.106	0.589	0.120	0.578	0.327
mGIC	0.558	0.100	0.565	0.124	0.574	0.327
GS^2	0.753	0.014	0.756	0.016	0.768	0.017
random proteins	100% similarity	σ				
Agreement	0.120	0.018				
simUI	0.391	0.029				
simGIC	0.176	0.029				
mUI	0.267	0.405				
mGIC	0.267	0.405				
GS^2	0.790	0.017				

Table A.4: *Average similarity results (as measured with the Agreement, simUI, simGIC, mUI, mGIC and GS^2 metrics) and respective standard deviations (σ) for the PL4 family and derived sets with progressive (10% increments; 100 iterations) amounts of original PL4 proteins being replaced by random proteins taken from other CAZy families.*

PL5	Agreement	simUI	simGIC	mUI	mGIC	GS^2
similarity	1.000	1.000	1.000	1.000	1.000	1.000
random proteins	10%		20%		30%	
	similarity	σ	similarity	σ	similarity	σ
Agreement	0.403	0.110	0.269	0.062	0.226	0.055
simUI	0.869	0.006	0.718	0.010	0.622	0.015
simGIC	0.837	0.003	0.646	0.005	0.524	0.010
mUI	0.995	0.019	0.988	0.028	0.980	0.039
mGIC	0.995	0.019	0.988	0.028	0.980	0.039
GS^2	0.944	0.007	0.882	0.011	0.842	0.013
random proteins	40%		50%		60%	
	similarity	σ	similarity	σ	similarity	σ
Agreement	0.194	0.040	0.179	0.036	0.157	0.030
simUI	0.517	0.015	0.456	0.017	0.394	0.019
simGIC	0.389	0.012	0.306	0.012	0.221	0.015
mUI	0.973	0.050	0.957	0.073	0.905	0.122
mGIC	0.973	0.050	0.957	0.073	0.905	0.122
GS^2	0.801	0.014	0.781	0.015	0.760	0.016
random proteins	70%		80%		90%	
	similarity	σ	similarity	σ	similarity	σ
Agreement	0.148	0.026	0.137	0.021	0.131	0.020
simUI	0.368	0.022	0.350	0.026	0.358	0.029
simGIC	0.182	0.019	0.152	0.023	0.147	0.023
mUI	0.863	0.149	0.796	0.201	0.722	0.264
mGIC	0.863	0.149	0.796	0.201	0.722	0.264
GS^2	0.755	0.016	0.754	0.018	0.766	0.018
random proteins	100%					
	similarity	σ				
Agreement	0.126	0.019				
simUI	0.386	0.028				
simGIC	0.171	0.027				
mUI	0.199	0.353				
mGIC	0.198	0.353				
GS^2	0.787	0.021				

Table A.5: *Average similarity results (as measured with the Agreement, simUI, simGIC, mUI, mGIC and GS^2 metrics) and respective standard deviations (σ) for the PL5 family and derived sets with progressive (10% increments; 100 iterations) amounts of original PL5 proteins being replaced by random proteins taken from other CAZy families.*

PL6	Agreement	simUI	simGIC	mUI	mGIC	GS^2
similarity	0.579	0.752	0.568	0.489	0.489	0.953

random proteins	10%		20%		30%	
	similarity	σ	similarity	σ	similarity	σ
Agreement	0.323	0.079	0.253	0.055	0.220	0.050
simUI	0.662	0.022	0.587	0.026	0.514	0.027
simGIC	0.465	0.024	0.380	0.029	0.298	0.026
mUI	0.485	0.037	0.492	0.061	0.476	0.069
mGIC	0.485	0.037	0.492	0.061	0.476	0.069
GS^2	0.914	0.012	0.879	0.016	0.847	0.018

random proteins	40%		50%		60%	
	similarity	σ	similarity	σ	similarity	σ
Agreement	0.206	0.038	0.185	0.033	0.184	0.034
simUI	0.464	0.029	0.412	0.030	0.388	0.035
simGIC	0.238	0.029	0.190	0.027	0.160	0.027
mUI	0.468	0.097	0.484	0.142	0.500	0.187
mGIC	0.468	0.097	0.484	0.142	0.500	0.187
GS^2	0.826	0.017	0.800	0.019	0.790	0.023

random proteins	70%		80%		90%	
	similarity	σ	similarity	σ	similarity	σ
Agreement	0.178	0.038	0.176	0.031	0.173	0.033
simUI	0.367	0.039	0.366	0.040	0.370	0.041
simGIC	0.139	0.029	0.141	0.033	0.151	0.039
mUI	0.485	0.214	0.631	0.296	0.197	0.337
mGIC	0.484	0.215	0.630	0.297	0.197	0.337
GS^2	0.779	0.026	0.778	0.026	0.780	0.025

random proteins	100%	
	similarity	σ
Agreement	0.184	0.033
simUI	0.390	0.044
simGIC	0.176	0.044
mUI	0.162	0.335
mGIC	0.156	0.331
GS^2	0.788	0.025

Table A.6: *Average similarity results (as measured with the Agreement, simUI, simGIC, mUI, mGIC and GS^2 metrics) and respective standard deviations (σ) for the PL6 family and derived sets with progressive (10% increments; 100 iterations) amounts of original PL6 proteins being replaced by random proteins taken from other CAZy families.*

PL7	Agreement	simUI	simGIC	mUI	mGIC	GS^2
similarity	0.290	0.809	0.691	0.717	0.717	0.906
random proteins	10%		20%		30%	
	similarity	σ	similarity	σ	similarity	σ
Agreement	0.147	0.028	0.124	0.025	0.105	0.018
simUI	0.715	0.013	0.637	0.021	0.554	0.018
simGIC	0.571	0.018	0.471	0.025	0.369	0.020
mUI	0.713	0.022	0.713	0.036	0.709	0.052
mGIC	0.713	0.022	0.713	0.036	0.709	0.052
GS^2	0.873	0.010	0.850	0.020	0.825	0.020
random proteins	40%		50%		60%	
	similarity	σ	similarity	σ	similarity	σ
Agreement	0.102	0.017	0.095	0.015	0.093	0.013
simUI	0.498	0.020	0.445	0.019	0.410	0.019
simGIC	0.297	0.022	0.229	0.019	0.188	0.016
mUI	0.707	0.065	0.690	0.079	0.674	0.100
mGIC	0.707	0.065	0.690	0.079	0.674	0.100
GS^2	0.806	0.020	0.792	0.023	0.782	0.021
random proteins	70%		80%		90%	
	similarity	σ	similarity	σ	similarity	σ
Agreement	0.093	0.012	0.089	0.013	0.090	0.009
simUI	0.390	0.019	0.377	0.021	0.378	0.019
simGIC	0.162	0.016	0.149	0.017	0.154	0.015
mUI	0.675	0.135	0.668	0.186	0.622	0.251
mGIC	0.675	0.135	0.667	0.186	0.621	0.251
GS^2	0.778	0.019	0.779	0.018	0.780	0.018
random proteins	100%					
	similarity	σ				
Agreement	0.090	0.009				
simUI	0.390	0.023				
simGIC	0.175	0.022				
mUI	0.216	0.351				
mGIC	0.212	0.347				
GS^2	0.789	0.014				

Table A.7: *Average similarity results (as measured with the Agreement, simUI, simGIC, mUI, mGIC and GS^2 metrics) and respective standard deviations (σ) for the PL7 family and derived sets with progressive (10% increments; 100 iterations) amounts of original PL7 proteins being replaced by random proteins taken from other CAZy families.*

PL8	Agreement	simUI	simGIC	mUI	mGIC	GS^2
similarity	0.232	0.897	0.810	0.693	0.649	0.982

random proteins	10%		20%		30%	
	similarity	σ	similarity	σ	similarity	σ
Agreement	0.131	0.018	0.103	0.014	0.089	0.011
simUI	0.769	0.004	0.658	0.006	0.564	0.006
simGIC	0.663	0.005	0.536	0.007	0.426	0.007
mUI	0.689	0.008	0.683	0.016	0.677	0.023
mGIC	0.644	0.010	0.638	0.016	0.630	0.022
GS^2	0.926	0.003	0.877	0.004	0.837	0.005

random proteins	40%		50%		60%	
	similarity	σ	similarity	σ	similarity	σ
Agreement	0.078	0.009	0.069	0.007	0.064	0.006
simUI	0.487	0.007	0.424	0.007	0.382	0.007
simGIC	0.334	0.007	0.257	0.006	0.203	0.007
mUI	0.665	0.034	0.647	0.035	0.632	0.040
mGIC	0.619	0.031	0.600	0.031	0.582	0.040
GS^2	0.805	0.005	0.781	0.005	0.765	0.006

random proteins	70%		80%		90%	
	similarity	σ	similarity	σ	similarity	σ
Agreement	0.060	0.007	0.055	0.005	0.054	0.005
simUI	0.358	0.009	0.349	0.010	0.361	0.010
simGIC	0.168	0.008	0.150	0.008	0.152	0.010
mUI	0.616	0.045	0.572	0.063	0.534	0.085
mGIC	0.562	0.049	0.517	0.061	0.482	0.081
GS^2	0.758	0.007	0.758	0.007	0.770	0.007

random proteins	100%	
	similarity	σ
Agreement	0.052	0.005
simUI	0.387	0.013
simGIC	0.173	0.012
mUI	0.365	0.379
mGIC	0.364	0.379
GS^2	0.787	0.008

Table A.8: *Average similarity results (as measured with the Agreement, simUI, simGIC, mUI, mGIC and GS^2 metrics) and respective standard deviations (σ) for the PL8 family and derived sets with progressive (10% increments; 100 iterations) amounts of original PL8 proteins being replaced by random proteins taken from other CAZy families.*

PL9	Agreement	simUI	simGIC	mUI	mGIC	GS^2
similarity	0.132	0.564	0.409	0.451	0.451	0.848
random proteins	10%		20%		30%	
	similarity	σ	similarity	σ	similarity	σ
Agreement	0.104	0.008	0.093	0.011	0.086	0.011
simUI	0.510	0.014	0.465	0.017	0.426	0.017
simGIC	0.339	0.015	0.282	0.017	0.232	0.016
mUI	0.442	0.023	0.445	0.036	0.440	0.048
mGIC	0.443	0.023	0.445	0.036	0.440	0.048
GS^2	0.826	0.007	0.808	0.010	0.793	0.011
random proteins	40%		50%		60%	
	similarity	σ	similarity	σ	similarity	σ
Agreement	0.081	0.010	0.077	0.009	0.076	0.009
simUI	0.399	0.020	0.372	0.021	0.361	0.020
simGIC	0.196	0.018	0.163	0.017	0.145	0.015
mUI	0.448	0.061	0.445	0.079	0.454	0.128
mGIC	0.448	0.061	0.446	0.079	0.454	0.128
GS^2	0.783	0.013	0.772	0.014	0.769	0.014
random proteins	70%		80%		90%	
	similarity	σ	similarity	σ	similarity	σ
Agreement	0.075	0.008	0.075	0.008	0.073	0.008
simUI	0.359	0.018	0.357	0.017	0.368	0.018
simGIC	0.139	0.014	0.138	0.013	0.150	0.014
mUI	0.479	0.117	0.448	0.164	0.559	0.241
mGIC	0.478	0.117	0.448	0.164	0.558	0.242
GS^2	0.771	0.013	0.770	0.013	0.777	0.013
random proteins	100%					
	similarity	σ				
Agreement	0.078	0.010				
simUI	0.386	0.021				
simGIC	0.172	0.021				
mUI	0.191	0.318				
mGIC	0.186	0.311				
GS^2	0.787	0.013				

Table A.9: *Average similarity results (as measured with the Agreement, simUI, simGIC, mUI, mGIC and GS^2 metrics) and respective standard deviations (σ) for the PL9 family and derived sets with progressive (10% increments; 100 iterations) amounts of original PL9 proteins being replaced by random proteins taken from other CAZy families.*

PL10	Agreement	simUI	simGIC	mUI	mGIC	GS^2
similarity	0.261	0.742	0.633	0.753	0.713	0.931

random proteins	10%		20%		30%	
	similarity	σ	similarity	σ	similarity	σ
Agreement	0.154	0.029	0.124	0.022	0.112	0.019
simUI	0.663	0.013	0.591	0.017	0.535	0.018
simGIC	0.527	0.016	0.431	0.021	0.355	0.020
mUI	0.759	0.032	0.762	0.051	0.783	0.056
mGIC	0.724	0.040	0.733	0.062	0.761	0.065
GS^2	0.901	0.006	0.874	0.008	0.854	0.008

random proteins	40%		50%		60%	
	similarity	σ	similarity	σ	similarity	σ
Agreement	0.101	0.015	0.092	0.012	0.087	0.010
simUI	0.477	0.018	0.435	0.019	0.405	0.016
simGIC	0.278	0.019	0.224	0.017	0.187	0.013
mUI	0.782	0.091	0.790	0.123	0.796	0.131
mGIC	0.764	0.099	0.776	0.133	0.787	0.138
GS^2	0.831	0.010	0.814	0.010	0.801	0.010

random proteins	70%		80%		90%	
	similarity	σ	similarity	σ	similarity	σ
Agreement	0.085	0.010	0.085	0.010	0.083	0.009
simUI	0.386	0.016	0.375	0.020	0.374	0.019
simGIC	0.162	0.014	0.149	0.016	0.152	0.017
mUI	0.818	0.159	0.817	0.177	0.698	0.209
mGIC	0.813	0.165	0.814	0.179	0.695	0.210
GS^2	0.793	0.011	0.789	0.012	0.786	0.012

random proteins	100%	
	similarity	σ
Agreement	0.084	0.009
simUI	0.389	0.021
simGIC	0.174	0.020
mUI	0.254	0.364
mGIC	0.254	0.363
GS^2	0.789	0.014

Table A.10: *Average similarity results (as measured with the Agreement, simUI, simGIC, mUI, mGIC and GS^2 metrics) and respective standard deviations (σ) for the PL10 family and derived sets with progressive (10% increments; 100 iterations) amounts of original PL10 proteins being replaced by random proteins taken from other CAZy families.*

PL11	Agreement	simUI	simGIC	mUI	mGIC	GS^2
similarity	0.345	0.489	0.278	0.507	0.495	0.841
random proteins	10%		20%		30%	
	similarity	σ	similarity	σ	similarity	σ
Agreement	0.249	0.048	0.208	0.033	0.182	0.032
simUI	0.465	0.012	0.446	0.017	0.426	0.017
simGIC	0.249	0.011	0.225	0.015	0.200	0.015
mUI	0.496	0.033	0.490	0.033	0.491	0.037
mGIC	0.482	0.034	0.474	0.035	0.476	0.036
GS^2	0.831	0.007	0.822	0.011	0.812	0.011
random proteins	40%		50%		60%	
	similarity	σ	similarity	σ	similarity	σ
Agreement	0.163	0.029	0.153	0.028	0.139	0.023
simUI	0.412	0.024	0.401	0.025	0.392	0.025
simGIC	0.185	0.021	0.173	0.022	0.164	0.021
mUI	0.465	0.053	0.478	0.095	0.485	0.139
mGIC	0.453	0.050	0.464	0.092	0.476	0.139
GS^2	0.806	0.015	0.801	0.016	0.795	0.017
random proteins	70%		80%		90%	
	similarity	σ	similarity	σ	similarity	σ
Agreement	0.135	0.023	0.125	0.020	0.120	0.016
simUI	0.394	0.025	0.381	0.024	0.384	0.027
simGIC	0.166	0.025	0.156	0.021	0.165	0.024
mUI	0.560	0.246	0.547	0.351	0.369	0.401
mGIC	0.554	0.246	0.545	0.351	0.368	0.401
GS^2	0.796	0.015	0.788	0.016	0.788	0.019

random proteins	100%	
	similarity	σ
Agreement	0.116	0.018
simUI	0.386	0.029
simGIC	0.172	0.026
mUI	0.182	0.325
mGIC	0.181	0.324
GS^2	0.787	0.018

Table A.11: *Average similarity results (as measured with the Agreement, simUI, simGIC, mUI, mGIC and GS^2 metrics) and respective standard deviations (σ) for the PL11 family and derived sets with progressive (10% increments; 100 iterations) amounts of original PL11 proteins being replaced by random proteins taken from other CAZy families.*

PL12	Agreement	simUI	simGIC	mUI	mGIC	GS^2
similarity	0.859	0.996	0.988	0.985	0.981	1.000

random proteins	10%		20%		30%	
	similarity	σ	similarity	σ	similarity	σ
Agreement	0.284	0.061	0.197	0.047	0.159	0.032
simUI	0.865	0.006	0.729	0.007	0.615	0.009
simGIC	0.824	0.005	0.656	0.006	0.511	0.007
mUI	0.982	0.011	0.979	0.018	0.971	0.029
mGIC	0.978	0.012	0.975	0.019	0.966	0.029
GS^2	0.944	0.005	0.888	0.007	0.842	0.009

random proteins	40%		50%		60%	
	similarity	σ	similarity	σ	similarity	σ
Agreement	0.135	0.023	0.119	0.018	0.111	0.017
simUI	0.522	0.010	0.447	0.011	0.403	0.012
simGIC	0.389	0.007	0.294	0.010	0.230	0.010
mUI	0.965	0.035	0.955	0.044	0.932	0.078
mGIC	0.961	0.036	0.950	0.045	0.928	0.078
GS^2	0.805	0.010	0.776	0.010	0.763	0.011

random proteins	70%		80%		90%	
	similarity	σ	similarity	σ	similarity	σ
Agreement	0.098	0.013	0.094	0.011	0.088	0.011
simUI	0.366	0.015	0.352	0.016	0.359	0.019
simGIC	0.179	0.012	0.152	0.014	0.150	0.016
mUI	0.879	0.111	0.849	0.145	0.719	0.212
mGIC	0.874	0.112	0.846	0.146	0.715	0.212
GS^2	0.753	0.012	0.755	0.012	0.766	0.013

random proteins	100%	
	similarity	σ
Agreement	0.090	0.012
simUI	0.392	0.020
simGIC	0.176	0.019
mUI	0.241	0.366
mGIC	0.240	0.365
GS^2	0.790	0.014

Table A.12: *Average similarity results (as measured with the Agreement, simUI, simGIC, mUI, mGIC and GS^2 metrics) and respective standard deviations (σ) for the PL12 family and derived sets with progressive (10% increments; 100 iterations) amounts of original PL12 proteins being replaced by random proteins taken from other CAZy families.*

PL16	Agreement	simUI	simGIC	mUI	mGIC	GS^2
similarity	1.000	1.000	1.000	1.000	1.000	1.000
random proteins	10%		20%		30%	
	similarity	σ	similarity	σ	similarity	σ
Agreement	0.602	0.145	0.468	0.119	0.370	0.106
simUI	0.892	0.023	0.794	0.031	0.704	0.039
simGIC	0.848	0.016	0.710	0.023	0.587	0.027
mUI	0.991	0.027	0.989	0.035	0.973	0.053
mGIC	0.991	0.027	0.989	0.035	0.973	0.053
GS^2	0.965	0.017	0.933	0.023	0.900	0.029
random proteins	40%		50%		60%	
	similarity	σ	similarity	σ	similarity	σ
Agreement	0.329	0.087	0.253	0.054	0.232	0.045
simUI	0.638	0.040	0.529	0.040	0.484	0.041
simGIC	0.486	0.030	0.348	0.030	0.285	0.032
mUI	0.961	0.080	0.908	0.125	0.884	0.143
mGIC	0.961	0.080	0.908	0.125	0.884	0.143
GS^2	0.883	0.029	0.841	0.031	0.828	0.029
random proteins	70%		80%		90%	
	similarity	σ	similarity	σ	similarity	σ
Agreement	0.209	0.044	0.191	0.032	0.177	0.028
simUI	0.443	0.041	0.411	0.037	0.395	0.038
simGIC	0.233	0.033	0.193	0.031	0.176	0.032
mUI	0.831	0.198	0.780	0.233	0.631	0.308
mGIC	0.831	0.198	0.780	0.233	0.631	0.308
GS^2	0.812	0.028	0.800	0.025	0.795	0.026

random proteins	100%	
	similarity	σ
Agreement	0.169	0.029
simUI	0.387	0.040
simGIC	0.171	0.036
mUI	0.097	0.239
mGIC	0.095	0.234
GS^2	0.789	0.027

Table A.13: *Average similarity results (as measured with the Agreement, simUI, simGIC, mUI, mGIC and GS^2 metrics) and respective standard deviations (σ) for the PL16 family and derived sets with progressive (10% increments; 100 iterations) amounts of original PL16 proteins being replaced by random proteins taken from other CAZy families.*

PL17	Agreement	simUI	simGIC	mUI	mGIC	GS^2
similarity	1.000	1.000	1.000	1.000	1.000	1.000
random proteins	10%		20%		30%	
	similarity	σ	similarity	σ	similarity	σ
Agreement	0.425	0.087	0.292	0.067	0.244	0.052
simUI	0.864	0.009	0.741	0.010	0.636	0.011
simGIC	0.829	0.009	0.675	0.006	0.540	0.007
mUI	0.997	0.014	0.991	0.023	0.984	0.035
mGIC	0.997	0.014	0.991	0.023	0.984	0.035
GS^2	0.943	0.007	0.891	0.010	0.849	0.011
random proteins	40%		50%		60%	
	similarity	σ	similarity	σ	similarity	σ
Agreement	0.198	0.040	0.182	0.032	0.164	0.027
simUI	0.522	0.015	0.453	0.016	0.405	0.020
simGIC	0.393	0.009	0.304	0.010	0.235	0.014
mUI	0.953	0.066	0.954	0.078	0.938	0.098
mGIC	0.953	0.066	0.954	0.078	0.938	0.098
GS^2	0.805	0.014	0.777	0.018	0.763	0.017
random proteins	70%		80%		90%	
	similarity	σ	similarity	σ	similarity	σ
Agreement	0.152	0.025	0.141	0.024	0.133	0.021
simUI	0.363	0.020	0.351	0.028	0.356	0.027
simGIC	0.175	0.018	0.153	0.023	0.147	0.025
mUI	0.903	0.129	0.794	0.202	0.698	0.281
mGIC	0.903	0.129	0.794	0.202	0.698	0.281
GS^2	0.752	0.015	0.752	0.021	0.763	0.019
random proteins	100%					
	similarity	σ				
Agreement	0.131	0.021				
simUI	0.385	0.031				
simGIC	0.169	0.030				
mUI	0.169	0.330				
mGIC	0.169	0.330				
GS^2	0.788	0.019				

Table A.14: *Average similarity results (as measured with the Agreement, simUI, simGIC, mUI, mGIC and GS^2 metrics) and respective standard deviations (σ) for the PL17 family and derived sets with progressive (10% increments; 100 iterations) amounts of original PL17 proteins being replaced by random proteins taken from other CAZy families.*

PL22	Agreement	simUI	simGIC	mUI	mGIC	GS^2
similarity	0.541	0.791	0.659	0.621	0.621	0.95

random proteins	10%		20%		30%	
	similarity	σ	similarity	σ	similarity	σ
Agreement	0.340	0.057	0.267	0.050	0.228	0.045
simUI	0.719	0.015	0.619	0.018	0.532	0.020
simGIC	0.577	0.021	0.459	0.025	0.356	0.025
mUI	0.625	0.027	0.612	0.037	0.596	0.054
mGIC	0.625	0.027	0.612	0.037	0.596	0.054
GS^2	0.916	0.007	0.871	0.011	0.832	0.014

random proteins	40%		50%		60%	
	similarity	σ	similarity	σ	similarity	σ
Agreement	0.200	0.040	0.179	0.029	0.173	0.037
simUI	0.463	0.023	0.413	0.021	0.379	0.024
simGIC	0.280	0.025	0.220	0.024	0.177	0.023
mUI	0.595	0.105	0.615	0.147	0.629	0.203
mGIC	0.595	0.105	0.615	0.147	0.629	0.203
GS^2	0.801	0.017	0.778	0.018	0.767	0.018

random proteins	70%		80%		90%	
	similarity	σ	similarity	σ	similarity	σ
Agreement	0.158	0.026	0.143	0.021	0.141	0.021
simUI	0.356	0.026	0.348	0.023	0.362	0.031
simGIC	0.150	0.023	0.138	0.022	0.147	0.026
mUI	0.664	0.257	0.632	0.272	0.524	0.351
mGIC	0.664	0.257	0.632	0.272	0.524	0.351
GS^2	0.759	0.02	0.760	0.017	0.771	0.023

random proteins	100%	
	similarity	σ
Agreement	0.146	0.025
simUI	0.387	0.031
simGIC	0.171	0.029
mUI	0.170	0.313
mGIC	0.163	0.306
GS^2	0.788	0.021

Table A.15: *Average similarity results (as measured with the Agreement, simUI, simGIC, mUI, mGIC and GS^2 metrics) and respective standard deviations (σ) for the PL22 family and derived sets with progressive (10% increments; 100 iterations) amounts of original PL22 proteins being replaced by random proteins taken from other CAZy families.*

Appendix B

Completeness results for randomization assays

Each tested CAZy PL families (PL1-PL12, PL16, PL17 and PL22) were altered by progressively replacing discrete amounts (with 10% increments) of original family proteins with the same amounts of proteins randomly selected from the CAZy database. Each of these created sets was measured with the Leaf-completeness and IC-completeness metrics.

% random proteins	Leaf-completeness		IC-completeness	
	%	σ	%	σ
0	40.48	-	6.88	-
10	39.06	1.35	6.43	0.42
20	40.23	1.87	5.86	0.62
30	42.14	2.11	5.29	0.63
40	41.94	3.59	4.90	0.79
50	42.61	4.57	4.30	0.74
60	43.08	5.33	3.74	0.86
70	45.05	5.64	3.37	0.91
80	44.61	7.21	2.90	0.78
90	44.07	7.79	2.17	0.79
100	44.80	8.91	1.76	0.63

Table B.1: *Average completeness results (as measured with the Leaf-completeness and IC-completness [IC threshold = 0.7] metrics) and the respective standard deviations (σ) for the PL1 family and derived sets with progressive (10% increments; 100 iterations) amounts of original PL1 proteins being replaced by random proteins taken from other CAZy families.*

% random proteins	Leaf-completeness		IC-completeness	
	%	σ	%	σ
0	61.76	-	29.41	-
10	62.91	3.16	26.85	2.48
20	63.03	4.73	24.32	3.30
30	61.38	5.98	20.68	3.59
40	61.68	6.20	18.35	4.16
50	60.41	7.10	15.26	4.47
60	59.38	6.74	13.94	4.09
70	59.82	8.24	11.32	4.25
80	59.56	8.20	7.56	3.63
90	59.15	9.10	5.26	3.06
100	57.32	8.58	1.65	2.13

Table B.2: *Average completeness results (as measured with the Leaf-completeness and IC-completness [IC threshold = 0.7] metrics) and the respective standard deviations (σ) for the PL2 family and derived sets with progressive (10% increments; 100 iterations) amounts of original PL2 proteins being replaced by random proteins taken from other CAZy families.*

% random proteins	Leaf-completeness		IC-completeness	
	%	σ	%	σ
0	100	-	0.9	-
10	95.93	1.15	1.04	0.37
20	91.08	1.81	1.02	0.44
30	86.18	2.52	1.15	0.55
40	81.34	2.91	1.32	0.64
50	75.75	3.70	1.32	0.66
60	70.96	4.32	1.41	0.86
70	66.15	5.55	1.48	0.83
80	59.81	6.20	1.62	0.77
90	54.22	7.72	1.67	0.87
100	49.10	8.39	1.86	0.87

Table B.3: *Average completeness results (as measured with the Leaf-completeness and IC-completness [IC threshold = 0.7] metrics) and the respective standard deviations (σ) for the PL3 family and derived sets with progressive (10% increments; 100 iterations) amounts of original PL3 proteins being replaced by random proteins taken from other CAZy families.*

% random proteins	Leaf-completeness		IC-completeness	
	%	σ	%	σ
0	100	-	0	-
10	92.61	16.55	0.27	0.76
20	89.20	13.10	0.41	1.15
30	83.10	15.91	0.44	0.94
40	76.54	16.54	0.80	1.54
50	73.20	15.07	0.85	1.63
60	67.32	15.02	1.00	1.62
70	65.90	13.00	1.10	1.48
80	64.05	10.03	1.02	1.59
90	59.66	9.65	1.71	2.00
100	57.90	8.98	1.44	2.01

Table B.4: *Average completeness results (as measured with the Leaf-completeness and IC-completness [IC threshold = 0.7] metrics) and the respective standard deviations (σ) for the PL4 family and derived sets with progressive (10% increments; 100 iterations) amounts of original PL4 proteins being replaced by random proteins taken from other CAZy families.*

% random proteins	Leaf-completeness		IC-completeness	
	%	σ	%	σ
0	100	-	0	-
10	98.40	1.91	0.11	0.56
20	93.23	3.49	0.26	0.82
30	90.51	4.18	0.57	1.21
40	85.80	5.07	0.51	1.10
50	81.11	5.33	0.57	1.34
60	76.26	6.12	1.11	1.80
70	71.97	7.11	0.97	1.52
80	67.20	7.56	1.54	2.34
90	62.49	7.29	1.54	2.08
100	58.34	8.84	1.77	2.05

Table B.5: *Average completeness results (as measured with the Leaf-completeness and IC-completness [IC threshold = 0.7] metrics) and the respective standard deviations (σ) for the PL5 family and derived sets with progressive (10% increments; 100 iterations) amounts of original PL5 proteins being replaced by random proteins taken from other CAZy families.*

% random proteins	Leaf-completeness		IC-completeness	
	%	σ	%	σ
0	35	-	5	-
10	40.40	4.04	4.60	1.69
20	42.55	6.22	4.00	2.45
30	44.65	6.57	4.20	2.98
40	46.95	7.74	3.20	3.28
50	51.00	9.19	3.55	3.56
60	51.95	10.70	2.90	3.55
70	55.15	10.38	2.75	3.11
80	59.60	10.19	2.45	3.43
90	62.75	10.89	2.30	3.12
100	61.10	11.01	1.40	2.75

Table B.6: *Average completeness results (as measured with the Leaf-completeness and IC-completness [IC threshold = 0.7] metrics) and the respective standard deviations (σ) for the PL6 family and derived sets with progressive (10% increments; 100 iterations) amounts of original PL6 proteins being replaced by random proteins taken from other CAZy families.*

% random proteins	Leaf-completeness		IC-completeness	
	%	σ	%	σ
0	17.19	-	3.12	-
10	22.19	2.05	3.09	0.73
20	25.47	3.58	2.72	1.05
30	30.03	3.54	2.64	1.18
40	33.38	4.15	2.38	1.37
50	37.28	4.65	2.47	1.42
60	41.62	5.11	2.25	1.52
70	45.92	5.82	2.14	1.66
80	49.19	8.16	2.39	1.94
90	51.69	8.31	1.81	1.78
100	55.39	7.78	1.95	1.72

Table B.7: *Average completeness results (as measured with the Leaf-completeness and IC-completness [IC threshold = 0.7] metrics) and the respective standard deviations (σ) for the PL7 family and derived sets with progressive (10% increments; 100 iterations) amounts of original PL7 proteins being replaced by random proteins taken from other CAZy families.*

% random proteins	Leaf-completeness		IC-completeness	
	%	σ	%	σ
0	26.37	-	3.85	-
10	34.16	17.67	3.68	0.45
20	37.44	17.31	3.31	0.64
30	42.73	18.34	3.18	0.87
40	45.19	17.62	2.99	0.94
50	44.10	13.99	2.74	1.09
60	45.48	10.63	2.54	1.05
70	47.97	9.35	2.39	1.06
80	47.67	7.36	2.15	0.91
90	48.08	6.99	1.86	0.95
100	50.55	7.59	1.95	1.10

Table B.8: *Average completeness results (as measured with the Leaf-completeness and IC-completness [IC threshold = 0.7] metrics) and the respective standard deviations (σ) for the PL8 family and derived sets with progressive (10% increments; 100 iterations) amounts of original PL8 proteins being replaced by random proteins taken from other CAZy families.*

% random proteins	Leaf-completeness		IC-completeness	
	%	σ	%	σ
0	39.02	-	10.98	-
10	40.11	2.8	10.21	1.10
20	41.52	3.65	9.04	1.55
30	44.05	4.02	8.32	1.85
40	45.04	4.71	7.22	1.96
50	47.10	5.67	6.34	1.94
60	47.77	5.94	5.33	2.05
70	49.78	5.87	4.50	1.81
80	51.54	6.71	3.72	1.78
90	51.21	8.19	2.67	1.62
100	54.22	8.11	2.02	1.40

Table B.9: *Average completeness results (as measured with the Leaf-completeness and IC-completness [IC threshold = 0.7] metrics) and the respective standard deviations (σ) for the PL9 family and derived sets with progressive (10% increments; 100 iterations) amounts of original PL9 proteins being replaced by random proteins taken from other CAZy families.*

% random proteins	Leaf-completeness		IC-completeness	
	%	σ	%	σ
0	20.55	-	0	-
10	23.70	2.30	0.16	0.52
20	26.74	3.31	0.38	0.64
30	30.22	7.44	0.45	0.80
40	33.52	4.16	0.62	0.83
50	37.40	7.15	0.84	1.04
60	40.58	6.00	0.99	1.15
70	46.81	9.68	1.49	1.45
80	49.44	9.33	1.25	1.33
90	52.48	8.02	1.56	1.51
100	55.85	6.84	1.74	1.35

Table B.10: *Average completeness results (as measured with the Leaf-completeness and IC-completness [IC threshold = 0.7] metrics) and the respective standard deviations (σ) for the PL10 family and derived sets with progressive (10% increments; 100 iterations) amounts of original PL10 proteins being replaced by random proteins taken from other CAZy families.*

% random proteins	Leaf-completeness		IC-completeness	
	%	σ	%	σ
0	43.90	-	0	-
10	44.98	10.87	0.20	0.66
20	47.98	10.17	0.39	1.02
30	50.12	10.78	0.56	1.08
40	48.34	11.86	0.54	1.12
50	51.78	9.58	0.73	1.22
60	52.59	9.84	0.95	1.42
70	53.22	9.51	1.37	1.67
80	55.46	8.40	1.22	1.84
90	55.88	8.76	1.22	1.64
100	58.39	7.55	1.85	2.10

Table B.11: *Average completeness results (as measured with the Leaf-completeness and IC-completness [IC threshold = 0.7] metrics) and the respective standard deviations (σ) for the PL11 family and derived sets with progressive (10% increments; 100 iterations) amounts of original PL11 proteins being replaced by random proteins taken from other CAZy families.*

% random proteins	Leaf-completeness		IC-completeness	
	%	σ	%	σ
0	100	-	1.47	-
10	97.37	1.55	1.53	0.72
20	92.76	2.61	1.49	0.96
30	88.15	3.10	1.57	1.16
40	83.56	4.25	1.57	1.22
50	79.74	4.21	1.62	1.24
60	75.19	5.40	1.81	1.58
70	71.53	5.67	1.69	1.53
80	64.71	6.93	1.72	1.44
90	60.87	6.58	1.63	1.42
100	54.50	8.54	1.82	1.61

Table B.12: *Average completeness results (as measured with the Leaf-completeness and IC-completness [IC threshold = 0.7] metrics) and the respective standard deviations (σ) for the PL12 family and derived sets with progressive (10% increments; 100 iterations) amounts of original PL12 proteins being replaced by random proteins taken from other CAZy families.*

% random proteins	Leaf-completeness		IC-completeness	
	%	σ	%	σ
0	100	-	0	-
10	98.91	2.04	0.05	0.45
20	95.64	3.40	0.18	1.10
30	91.68	4.81	0.50	1.42
40	88.00	5.94	0.55	1.61
50	82.64	6.50	1.41	2.38
60	78.73	8.23	1.50	2.95
70	75.00	8.12	1.41	2.63
80	69.32	8.54	1.18	2.19
90	66.86	9.23	1.23	2.21
100	59.27	10.83	1.55	2.67

Table B.13: *Average completeness results (as measured with the Leaf-completeness and IC-completness [IC threshold = 0.7] metrics) and the respective standard deviations (σ) for the PL16 family and derived sets with progressive (10% increments; 100 iterations) amounts of original PL16 proteins being replaced by random proteins taken from other CAZy families.*

% random proteins	Leaf-completeness		IC-completeness	
	%	σ	%	σ
0	100	-	0	-
10	98.18	2.01	0.15	0.66
20	94.70	3.16	0.27	0.87
30	90.85	3.76	0.42	1.36
40	85.33	5.19	0.88	1.62
50	81.33	6.66	0.76	1.38
60	77.00	6.30	1.06	1.84
70	72.52	6.98	1.03	1.62
80	69.42	7.41	1.42	2.16
90	63.09	8.01	1.76	2.15
100	58.03	9.27	1.67	2.51

Table B.14: *Average completeness results (as measured with the Leaf-completeness and IC-completness [IC threshold = 0.7] metrics) and the respective standard deviations (σ) for the PL17 family and derived sets with progressive (10% increments; 100 iterations) amounts of original PL17 proteins being replaced by random proteins taken from other CAZy families.*

% random proteins	Leaf-completeness		IC-completeness	
	%	σ	%	σ
0	79.31	-	75.86	-
10	80.38	2.47	71.34	2.32
20	77.72	4.37	63.38	3.23
30	74.66	5.28	54.97	3.64
40	73.28	6.21	47.69	4.14
50	70.21	7.70	40.14	4.38
60	67.21	7.67	32.31	4.19
70	66.07	7.61	25.03	3.86
80	64.38	9.84	17.03	4.10
90	61.41	8.48	9.21	3.12
100	59.76	8.38	1.69	2.21

Table B.15: *Average completeness results (as measured with the Leaf-completeness and IC-completness [IC threshold = 0.7] metrics) and the respective standard deviations (σ) for the PL22 family and derived sets with progressive (10% increments; 100 iterations) amounts of original PL22 proteins being replaced by random proteins taken from other CAZy families.*

References

Abascal, F. & Valencia, A. (2003). Automatic Annotation of Protein Function Based on Family Identification. *Science And Technology*, **692**, 683– 692. 38, 44

Aerts, S., Haeussler, M., van Vooren, S., Griffith, O.L., Hulpiau, P., Jones, S.J., Montgomery, S.B., Bergman, C.M. & Consortium, T.O.R.A. (2008). Text-mining assisted regulatory annotation. *Genome Biology*, **9**, R31. 42, 46

Alexa, A., Rahnenführer, J. & Lengauer, T. (2006). Improved scoring of functional groups from gene expression data by decorrelating go graph structure. *Bioinformatics*, **22**, 1600–1607. 30

Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.*, **25**, 3389–3402. 19, 37

Andrade, M., Brown, N., Leroy, C., Hoersch, S., de Daruvar, A., Reich, C., Franchini, A., Tamames, J., Valencia, A., Ouzounis, C. & Sander, C. (1999). Automated genome sequence analysis and annotation. *Bioinformatics*, **15**, 391–412. 37, 43

Aparicio, G., Götz, S., Conesa, a., Segrelles, D., Blanquer, I., García, J.M., Hernandez, V., Robles, M. & Talon, M. (2006). Blast2GO goes grid: developing a grid-enabled prototype for functional genomics analysis. *Studies in health technology and informatics*, **120**, 194–204. 46

Attwood, T.K. (2003). PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Research*, **31**, 400–402. 18

Bailey, L.C., Fischer, S., Schug, J., Crabtree, J., Gibson, M. & Overton, G.C. (1998). GAIA: Framework Annotation of Genomic Sequence. *Genome Res.*, **8**, 234–250. 39

Bairoch, A. (2000). The ENZYME database in 2000. *Nucleic Acids Research*, **28**, 304–305. 22

Barrell, D., Dimmer, E., Huntley, R.P., Binns, D., O'Donovan, C. & Apweiler, R. (2009). The GOA database in 2009–an integrated Gene Ontology Annotation resource. *Nucleic acids research*, **37**, D396–403. 84

Bastos, H., Faria, D., Pesquita, C. & Falcão, A. (2007). Using go terms to evaluate protein clustering. In *BioOntologies SIG at ISMB/ECCB - 15th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB)*. 56

Bastos, H., Clarke, L.A. & Couto, F.M. (2013). Annotation extension through protein family annotation coherence metrics. *Frontiers in Genetics*, **4**. 6

Bastos, H.P., Sousa, L., Clarke, L.A. & Couto, F.M. (2015). Gryfun: A web application for go term annotation visualization and analysis in protein sets. *PLoS ONE*, **10**, e0119631. 6, 83

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. & Wheeler, D.L. (2008). GenBank. *Nucleic acids research*, **36**, D25–30. 17

Berman, H., Henrick, K., Nakamura, H. & Markley, J.L. (2007). The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic acids research*, **35**, D301–3. 17

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. & A (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL. *Nucleic Acids Res*, **31**, 365–370. 38

Booton, R. & Lindsay, M. (2014). Emerging role of microrna s and long non-coding rnas in respiratory disease. *Chest*, **146**, 193–204. 101

Bowers, P., Pellegrini, M., Thompson, M., Fierro, J., Yeates, T. & Eisenberg, D. (2004). Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biology*, **5**, R35. 37, 41

Bru, C., Courcelle, E., Carrère, S., Beausse, Y., Dalmar, S. & Kahn, D. (2005). The ProDom database of protein domain families: more emphasis on 3D. *Nucleic acids research*, **33**, D212–5. 19

Cantarel, B.L., Coutinho, P.M., Rancurel, C., Bernard, T., Lombard, V. & Henrissat, B. (2009). The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic acids research*, **37**, D233–8. 4, 20, 56

Chatziioannou, A.A. & Moulos, P. (2011). Exploiting statistical methodologies and controlled vocabularies for prioritized functional analysis of genomic experiments: The stranger web application. *Frontiers in Neuroscience*, **5**. 83

Chua, H.N., Sung, W.K. & Wong, L. (2006). Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. *Bioinformatics (Oxford, England)*, **22**, 1623–30. 37, 42

Clarke, L.A., Sousa, L., Barreto, C. & Amaral, M.D. (2013). Changes in transcriptome of native nasal epithelium expressing f508del-cftr and intersecting data from comparable studies. *Respiratory Research*, **14**, 38. xxviii, 96, 100

Cochrane, G., Akhtar, R., Bonfield, J., Bower, L., Demiralp, F., Faruque, N., Gibson, R., Hoad, G., Hubbard, T., Hunter, C., Jang, M., Juhos, S., Leinonen, R., Leonard, S., Lin, Q., Lopez, R., Lorenc, D., McWilliam, H., Mukherjee, G., Plaister, S., Radhakrishnan, R., Robinson, S., Sobhany, S., Hoopen, P.T., Vaughan, R., Zalunin, V. & Birney, E. (2009). Petabyte-scale innovations at the European Nucleotide Archive. *Nucl. Acids Res.*, **37**, D19–25. 16

Couto, F. & Silva, M. (2011). Disjunctive shared information between ontology concepts: application to gene ontology. *Journal of Biomedical Semantics*, **2**, 5. 26

Couto, F., Silva, M. & Coutinho, P. (2003). ProFAL: Protein functional annotation through literature. In *VII Conference on Software Engineering and Databases (JISBD)*, 747–756. 46

Couto, F., Silva, M., Lee, V., Dimmer, E., Camon, E., Apweiler, R., Kirsch, H. & Rebholz-Schuhmann, D. (2006a). Goannotator: linking protein go annotations to evidence text. *Journal of Biomedical Discovery and Collaboration*, **1**, 19. 42

Couto, F., Silva, M.J. & Coutinho, P.M. (2006b). Validating associations in biological databases. In *Proceedings of the 2006 ACM CIKM International Conference on Information and Knowledge Management, Arlington, Virginia, USA, November 6-11, 2006*, vol. November, 142–151. 46

Crick, F.H.C. (1958). On protein synthesis. *The Symposia of the Society for Experimental Biology*, **12**, 138–163. 11

del Val, C., Glatting, K.H. & Suhai, S. (2003). cDNA2Genome: a tool for mapping and annotating cDNAs. *BMC bioinformatics*, **4**, 39. 39

Deng, M., Zhang, K., Mehta, S., Chen, T. & Sun, F. (2002). Prediction of protein function using protein-protein interaction data. *Proceedings / IEEE Computer Society Bioinformatics Conference. IEEE Computer Society Bioinformatics Conference*, **1**, 197–206. 42

Devos, D. & Valencia, A. (2001). Intrinsic errors in genome annotation. *Trends in Genetics*, **17**, 429 – 431. 21

Diaz-Diaz, N. & Aguilar-Ruiz, J. (2011). Go-based functional dissimilarity of gene sets. *BMC Bioinformatics*, **12**, 360. 49, 66

Dimmer, E.C., Huntley, R.P., Alam-Faruque, Y., Sawford, T., O'Donovan, C., Martin, M.J., Bely, B., Browne, P., Mun Chan, W., Eberhardt, R., Gardner, M., Laiho, K., Legge, D., Magrane, M., Pichler, K., Poggioli, D., Sehra, H., Auchincloss, A., Axelsen, K., Blatter, M.C., Boutet, E., Braconi-Quintaje, S., Breuza, L., Bridge, A., Coudert, E., Estreicher, A., Famiglietti, L., Ferro-Rojas, S., Feuermann, M., Gos, A., Gruaz-Gumowski, N., Hinz, U., Hulo, C., James, J., Jimenez, S., Jungo, F., Keller, G., Lemercier, P., Lieberherr, D., Masson, P., Moinat, M., Pedruzzi, I., Poux, S., Rivoire, C., Roechert, B., Schneider, M., Stutz, A., Sundaram, S., Tognolli, M., Bougueleret, L., Argoud-Puy, G., Cusin, I., Duek-Roggli, P., Xenarios, I. & Apweiler, R. (2012). The uniprot-go annotation database in 2011. *Nucleic Acids Research*, **40**, D565–D570. 24

du Plessis, L., Skunca, N. & Dessimoz, C. (2011). The what, where, how and why of gene ontology—a primer for bioinformaticians. *Briefings in Bioinformatics*, **12**, 723–735. 25

Eden, E., Navon, R., Steinfeld, I., Lipson, D. & Yakhini, Z. (2009). Gorilla: a tool for discovery and visualization of enriched go terms in ranked gene lists. *BMC Bioinformatics*, **10**, 48. 83, 99

Ellson, J., Gansner, E.R., Koutsofios, E., North, S.C. & Woodhull, G. (2001). Graphviz - Open Source Graph Drawing Tools. *Graph Drawing*, 483–484. 84

Elsik, C.G., Worley, K.C., Zhang, L., Milshina, N.V., Jiang, H., Reese, J.T., Childs, K.L., Venkatraman, A., Dickens, C.M., Weinstock, G.M. & Gibbs, R.A. (2006). Community annotation: procedures, protocols, and supporting tools. *Genome research*, **16**, 1329–33. 33

Enault, F., Suhre, K. & Claverie, J.M. (2005). Phydbac "Gene Function Predictor": a gene annotation tool based on genomic context analysis. *BMC bioinformatics*, **6**, 247. 42

Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., Gavin, O.L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E.L.L., Eddy, S.R. & Bateman, A. (2010). The Pfam protein families database. *Nucleic acids research*, **38**, D211–22. 19

Finn, R.D., Clements, J. & Eddy, S.R. (2011). Hmmer web server: interactive sequence similarity searching. *Nucleic Acids Research*, **39**, 29–37. 5, 77

Fleischmann, A., Darsow, M., Degtyarenko, K., Fleischmann, W., Boyce, S., Axelsen, K.B., Bairoch, A., Schomburg, D., Tipton, K.F. & Apweiler, R. (2004). IntEnz, the integrated relational enzyme database. *Nucleic acids research*, **32**, D434–7. 22

Fleischmann, W., Moller, S., Gateau, A. & Apweiler, R. (1999). A novel method for automatic functional annotation of proteins. *Bioinformatics*, **15**, 228–233. 38, 43

Frangeul, L., Glaser, P., Rusniok, C., Buchrieser, C., Duchaud, E., Dehoux, P. & Kunst, F. (2004). CAAT-Box, Contigs-Assembly and Annotation Tool-Box for genome sequencing projects. *Bioinformatics (Oxford, England)*, **20**, 790–7. 39

Gansner, E.R. & North, S.C. (2000). An open graph visualization system and its applications to software engineering. *Software: Practice and Experience*, **30**, 1203–1233. 84

Gene Ontology Consortium, T. (2000). Gene ontology: Tool for the unification of biology. *Nature Genetics*, **25**, 25–29. 2, 22, 86

Gentleman, R. (2005). Visualizing and Distances Using GO. Tech. rep. 28, 68

Gibbs, R.a., Weinstock, G.M., Metzker, M.L., Muzny, D.M., Sodergren, E.J., Scherer, S., Scott, G., Steffen, D., Worley, K.C., Burch, P.E., Okwuonu, G., Hines, S., Lewis, L., DeRamo, C., Delgado, O., Dugan-Rocha, S., Miner, G., Morgan, M., Hawes, A., Gill, R., Celera, Holt, R.a., Adams, M.D., Amanatides, P.G.,

Baden-Tillson, H., Barnstead, M., Chin, S., Evans, C.a., Ferriera, S., Fosler, C., Glodek, A., Gu, Z., Jennings, D., Kraft, C.L., Nguyen, T., Pfannkoch, C.M., Sitter, C., Sutton, G.G., Venter, J.C., Woodage, T., Smith, D., Lee, H.M., Gustafson, E., Cahill, P., Kana, A., Doucette-Stamm, L., Weinstock, K., Fechtel, K., Weiss, R.B., Dunn, D.M., Green, E.D., Blakesley, R.W., Bouffard, G.G., De Jong, P.J., Osoegawa, K., Zhu, B., Marra, M., Schein, J., Bosdet, I., Fjell, C., Jones, S., Krzywinski, M., Mathewson, C., Siddiqui, A., Wye, N., McPherson, J., Zhao, S., Fraser, C.M., Shetty, J., Shatsman, S., Geer, K., Chen, Y., Abramzon, S., Nierman, W.C., Havlak, P.H., Chen, R., Durbin, K.J., Egan, A., Ren, Y., Song, X.Z., Li, B., Liu, Y., Qin, X., Cawley, S., Cooney, a.J., D'Souza, L.M., Martin, K., Wu, J.Q., Gonzalez-Garay, M.L., Jackson, A.R., Kalafus, K.J., McLeod, M.P., Milosavljevic, A., Virk, D., Volkov, A., Wheeler, D.a., Zhang, Z., Bailey, J.a., Eichler, E.E., Tuzun, E., Birney, E., Mongin, E., Ureta-Vidal, A., Woodwark, C., Zdobnov, E., Bork, P., Suyama, M., Torrents, D., Alexandersson, M., Trask, B.J., Young, J.M., Huang, H., Wang, H., Xing, H., Daniels, S., Gietzen, D., Schmidt, J., Stevens, K., Vitt, U., Wingrove, J., Camara, F., Mar Albà, M., Abril, J.F., Guigo, R., Smit, A., Dubchak, I., Rubin, E.M., Couronne, O., Poliakov, A., Hübner, N., Ganten, D., Goesele, C., Hummel, O., Kreitler, T., Lee, Y.A., Monti, J., Schulz, H., Zimdahl, H., Himmelbauer, H., Lehrach, H., Jacob, H.J., Bromberg, S., Gullings-Handley, J., Jensen-Seaman, M.I., Kwitek, A.E., Lazar, J., Pasko, D., Tonellato, P.J., Twigger, S., Ponting, C.P., Duarte, J.M., Rice, S., Goodstadt, L., Beatson, S.a., Emes, R.D., Winter, E.E., Webber, C., Brandt, P., Nyakatura, G., Adetobi, M., Chiaromonte, F., Elnitski, L., Eswara, P., Hardison, R.C., Hou, M., Kolbe, D., Makova, K., Miller, W., Nekrutenko, A., Riemer, C., Schwartz, S., Taylor, J., Yang, S., Zhang, Y., Lindpaintner, K., Andrews, T.D., Caccamo, M., Clamp, M., Clarke, L., Curwen, V., Durbin, R., Eyras, E., Searle, S.M., Cooper, G.M., Batzoglou, S., Brudno, M., Sidow, A., Stone, E.a., Payseur, B.a.,

Bourque, G., López-Otín, C., Puente, X.S., Chakrabarti, K., Chatterji, S., Dewey, C., Pachter, L., Bray, N., Yap, V.B., Caspi, A., Tesler, G., Pevzner, P.a., Haussler, D., Roskin, K.M., Baertsch, R., Clawson, H., Furey, T.S., Hinrichs, A.S., Karolchik, D., Kent, W.J., Rosenbloom, K.R., Trumbower, H., Weirauch, M., Cooper, D.N., Stenson, P.D., Ma, B., Brent, M., Arumugam, M., Shteynberg, D., Copley, R.R., Taylor, M.S., Riethman, H., Mudunuri, U., Peterson, J., Guyer, M., Felsenfeld, A., Old, S., Mockrin, S. & Collins, F. (2004). Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*, **428**, 493–521. 34

Gonzalez, O. & Zimmer, R. (2008). Assigning functional linkages to proteins using phylogenetic profiles and continuous phenotypes. *Bioinformatics (Oxford, England)*, **24**, 1257–63. 37

Götz, S., García-Gómez, J.M., Terol, J., Williams, T.D., Nagaraj, S.H., Nueda, M.J., Robles, M., Talón, M., Dopazo, J. & Conesa, A. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic acids research*, **36**, 3420–35. 38, 45

Griffith, O.L., Montgomery, S.B., Bernier, B., Chu, B., Kasaian, K., Aerts, S., Mahony, S., Sleumer, M.C., Bilenky, M., Haeussler, M., Griffith, M., Gallo, S.M., Giardine, B., Hooghe, B., Van Loo, P., Blanco, E., Ticoll, A., Lithwick, S., Portales-Casamar, E., Donaldson, I.J., Robertson, G., Wadelius, C., De Bleser, P., Vlieghe, D., Halfon, M.S., Wasserman, W., Hardison, R., Bergman, C.M. & Jones, S.J.M. (2008). ORegAnno: an open-access community-driven resource for regulatory annotation. *Nucleic acids research*, **36**, D107–13. 34

Gruber, T.R. (1993). A translation approach to portable ontology specifications. *Knowl. Acquis.*, **5**, 199–220. 25

Gruca, A., Sikora, M. & Polanski, A. (2011). Rulego: a logical rules-based tool for description of gene groups by means of gene ontology. *Nucleic Acids Research*, **39**, W293–W301. 49

Guo, X., Liu, R., Shriver, C.D., Hu, H. & Liebman, M.N. (2006). Assessing semantic similarity measures for the characterization of human regulatory pathways. *Bioinformatics*, **22**, 967–973. 29

Haft, D.H., Selengut, J.D., Richter, R.A., Harkins, D., Basu, M.K. & Beck, E. (2013). Tigrfams and genome properties in 2013. *Nucleic Acids Research*, **41**, D387–D395. 20

Hajj, R., Lesimple, P., Nawrocki-Raby, B., Birembaut, P., Puchelle, E. & Coraux, C. (2007). Human airway surface epithelial regeneration is delayed and abnormal in cystic fibrosis. *The Journal of Pathology*, **211**, 340–350. 101

Harris, N.L. (1997). Genotator: A Workbench for Sequence Annotation. *Genome Res.*, **7**, 754–762. 39

Henikoff, J.G. (2000). Increased coverage of protein families with the Blocks Database servers. *Nucleic Acids Research*, **28**, 228–230. 18

Hennig, S., Groth, D. & Lehrach, H. (2003). Automated Gene Ontology annotation for anonymous sequence data. *Nucleic Acids Research*, **31**, 3712–3715. 45

Huang, D.W., Sherman, B.T. & Lempicki, R.A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, **37**, 1–13. 30, 99

IUPAC-IUB Comm. on Biochem. Nomenclature (CBN) (1970). Abbreviations and symbols for nucleic acids, polynucleotides, and their constituents. *Biochemistry*, **9**, 4022–4027. 10

Jantzen, S., Sutherland, B., Minkley, D. & Koop, B. (2011). Go trimming: Systematically reducing redundancy in large gene ontology datasets. *BMC Research Notes*, **4**, 1–9. 51

Jensen, L., Gupta, R., Blom, N., Devos, D., Tamames, J., Kesmir, C., Nielsen, H., Stærfeldt, H., Rapacki, K., Workman, C., Andersen, C., Knudsen, S., Krogh, A., Valencia, A. & Brunak, S. (2002). Prediction of Human Protein Function from Post-translational Modifications and Localization Features. *Journal of Molecular Biology*, **319**, 1257–1265. 39, 45

Jensen, L.J., Gupta, R., Staerfeldt, H.H. & Brunak, S. (2003). Prediction of human protein function according to Gene Ontology categories. *Bioinformatics*, **19**, 635–642. 39, 45, 46

Jensen, L.J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M., Bork, P. & von Mering, C. (2009). STRING 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic acids research*, **37**, D412–6. 19

Jiang, J. & Conrath, D. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proc. of the Int'l. Conf. on Research in Computational Linguistics*, 19–33. 3, 27

Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M. & Hirakawa, M. (2006). From genomics to chemical genomics: new developments in KEGG. *Nucleic acids research*, **34**, D354–7. 22

Katoh, K. & Toh, H. (2008). Recent developments in the mafft multiple sequence alignment program. *Briefings in Bioinformatics*, **9**, 286–298. 5, 77

Kawai, J., Shinagawa, a., Shibata, K., Yoshino, M., Itoh, M., Ishii, Y., Arakawa, T., Hara, a., Fukunishi, Y., Konno, H., Adachi, J., Fukuda, S., Aizawa, K., Izawa, M., Nishi, K., Kiyosawa, H., Kondo, S., Yamanaka, I., Saito, T., Okazaki, Y., Gojobori, T., Bono, H., Kasukawa, T., Saito, R., Kadota, K., Matsuda, H., Ashburner, M., Batalov, S., Casavant, T., Fleischmann, W., Gaasterland,

T., Gissi, C., King, B., Kochiwa, H., Kuehl, P., Lewis, S., Matsuo, Y., Nikaido, I., Pesole, G., Quackenbush, J., Schriml, L.M., Staubli, F., Suzuki, R., Tomita, M., Wagner, L., Washio, T., Sakai, K., Okido, T., Furuno, M., Aono, H., Baldarelli, R., Barsh, G., Blake, J., Boffelli, D., Bojunga, N., Carninci, P., de Bonaldo, M.F., Brownstein, M.J., Bult, C., Fletcher, C., Fujita, M., Gariboldi, M., Gustincich, S., Hill, D., Hofmann, M., Hume, D.a., Kamiya, M., Lee, N.H., Lyons, P., Marchionni, L., Mashima, J., Mazzarelli, J., Mombaerts, P., Nordone, P., Ring, B., Ringwald, M., Rodriguez, I., Sakamoto, N., Sasaki, H., Sato, K., Schönbach, C., Seya, T., Shibata, Y., Storch, K.F., Suzuki, H., Toyo-oka, K., Wang, K.H., Weitz, C., Whittaker, C., Wilming, L., Wynshaw-Boris, a., Yoshida, K., Hasegawa, Y., Kawaji, H., Kohtsuki, S. & Hayashizaki, Y. (2001). Functional annotation of a full-length mouse cDNA collection. *Nature*, **409**, 685–90. 34

Khan, S., Situ, G., Decker, K. & Schmidt, C.J. (2003). GoFigure: Automated Gene OntologyTM annotation. *Bioinformatics*, **19**, 2484–2485. 45

Kretschmann, E., Fleischmann, W. & Apweiler, R. (2001). Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied on SWISS-PROT. *Bioinformatics*, **17**, 920–926. 38, 44

Lee, J., Katari, G. & Sachidanandam, R. (2005). Gobar: A gene ontology based analysis and visualization tool for gene sets. *BMC Bioinformatics*, **6**, 189. 83

Letovsky, S. & Kasif, S. (2003). Predicting protein function from protein/protein interaction data: a probabilistic approach. *Bioinformatics*, **19**, 197i–204. 42

Letunic, I., Doerks, T. & Bork, P. (2012). Smart 7: recent updates to the protein domain annotation resource. *Nucleic Acids Research*, **40**, D302–D305. 19

Lin, D. (1998). An information-theoretic definition of similarity. In *In Proceedings of the 15th International Conference on Machine Learning*, 296–304, Morgan Kaufmann. 3, 27

Lombard, V., Bernard, T., Rancurel, C., Brumer, H., Coutinho, P.M. & Henrissat, B. (2010). A hierarchical classification of polysaccharide lyases for glycogenomics. *Biochemical Journal*, **432**, 437–444. 96

Lord, P.W., Stevens, R.D., Brass, A. & Goble, C.A. (2003). Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation. *Bioinformatics*, **19**, 1275–1283. 2

Makarova, K., Sorokin, A., Novichkov, P., Wolf, Y. & Koonin, E. (2007). Clusters of orthologous genes for 41 archaeal genomes and implications for evolutionary genomics of archaea. *Biology Direct*, **2**, 33. 19

Martin, D.M.a., Berriman, M. & Barton, G.J. (2004). GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes. *BMC bioinformatics*, **5**, 178. 45

Mazumder, R., Natale, D.A., Julio, J.A.E., Yeh, L.S. & Wu, C.H. (2010). Community annotation in biology. *Biology direct*, **5**, 12. 34

McDermott, J. & Samudrala, R. (2003). Bioverse: functional, structural and contextual annotation of proteins and proteomes. *Nucleic Acids Research*, **31**, 3736–3737. 38, 47

Menda, N., Buels, R.M., Tecle, I. & Mueller, L.A. (2008). A community-based annotation framework for linking solanaceae genomes with phenomes. *Plant physiology*, **147**, 1788–99. 34

Meyer, F., Goesmann, A., McHardy, A.C., Bartels, D., Bekel, T., Clausen, J., Kalinowski, J., Linke, B., Rupp, O., Giegerich, R. & Puhler, A. (2003). GenDB–an open source genome annotation system for prokaryote genomes. *Nucleic Acids Research*, **31**, 2187–2195. 39

Miller, G.A. (1995). Wordnet: A lexical database for english. *Communications of the ACM*, **38**, 39–41. 3

Mural, R.J., Adams, M.D., Myers, E.W., Smith, H.O., Miklos, G.L.G., Wides, R., Halpern, A., Li, P.W., Sutton, G.G., Nadeau, J., Salzberg, S.L., Holt, R.A., Kodira, C.D., Lu, F., Chen, L., Deng, Z., Evangelista, C.C., Gan, W., Heiman, T.J., Li, J., Li, Z., Merkulov, G.V., Milshina, N.V., Naik, A.K., Qi, R., Shue, B.C., Wang, A., Wang, J., Wang, X., Yan, X., Ye, J., Yooseph, S., Zhao, Q., Zheng, L., Zhu, S.C., Biddick, K., Bolanos, R., Delcher, A.L., Dew, I.M., Fasulo, D., Flanigan, M.J., Huson, D.H., Kravitz, S.A., Miller, J.R., Mobarry, C.M., Reinert, K., Remington, K.A., Zhang, Q., Zheng, X.H., Nusskern, D.R., Lai, Z., Lei, Y., Zhong, W., Yao, A., Guan, P., Ji, R.R., Gu, Z., Wang, Z.Y., Zhong, F., Xiao, C., Chiang, C.C., Yandell, M., Wortman, J.R., Amanatides, P.G., Hladun, S.L., Pratts, E.C., Johnson, J.E., Dodson, K.L., Woodford, K.J., Evans, C.A., Gropman, B., Rusch, D.B., Venter, E., Wang, M., Smith, T.J., Houck, J.T., Tompkins, D.E., Haynes, C., Jacob, D., Chin, S.H., Allen, D.R., Dahlke, C.E., Sanders, R., Li, K., Liu, X., Levitsky, A.A., Majoros, W.H., Chen, Q., Xia, A.C., Lopez, J.R., Donnelly, M.T., Newman, M.H., Glodek, A., Kraft, C.L., Nodell, M., Ali, F., An, H.J., Baldwin-Pitts, D., Beeson, K.Y., Cai, S., Carnes, M., Carver, A., Caulk, P.M., Center, A., Chen, Y.H., Cheng, M.L., Coyne, M.D., Crowder, M., Danaher, S., Davenport, L.B., Desilets, R., Dietz, S.M., Doup, L., Dullaghan, P., Ferriera, S., Fosler, C.R., Gire, H.C., Gluecksmann, A., Gocayne, J.D., Gray, J., Hart, B., Haynes, J., Hoover, J., Howland, T., Ibegwam, C., Jalali, M., Johns, D., Kline, L., Ma, D.S., MacCawley, S., Magoon, A., Mann, F., May, D., McIntosh, T.C., Mehta, S., Moy, L., Moy, M.C., Murphy, B.J., Murphy, S.D., Nelson, K.A., Nuri, Z., Parker, K.A., Prudhomme, A.C., Puri, V.N., Qureshi, H., Raley, J.C., Reardon, M.S., Regier, M.A., Rogers, Y.H.C., Romblad, D.L., Schutz, J., Scott, J.L., Scott, R., Sitter, C.D., Smallwood, M.,

Sprague, A.C., Stewart, E., Strong, R.V., Suh, E., Sylvester, K., Thomas, R., Tint, N.N., Tsonis, C., Wang, G., Wang, G., Williams, M.S., Williams, S.M., Windsor, S.M., Wolfe, K., Wu, M.M., Zaveri, J., Chaturvedi, K., Gabrielian, A.E., Ke, Z., Sun, J., Subramanian, G. & Venter, J.C. (2002). A Comparison of Whole-Genome Shotgun-Derived Mouse Chromosome 16 and the Human Genome. *Science*, **296**, 1661–1671. 39

Nakamura, Y., Cochrane, G. & Karsch-Mizrachi, I. (2013). The international nucleotide sequence database collaboration. *Nucleic Acids Research*, **41**, D21–D24. 17

Nehrt, N.L., Clark, W.T., Radivojac, P. & Hahn, M.W. (2011). Testing the ortholog conjecture with comparative functional genomic data from mammals. *PLoS Comput Biol*, **7**, e1002073. 3

Nelson, D.L. & Cox, M.M. (2004). *Lehninger Principles of Biochemistry, Fourth Edition*. W. H. Freeman, fourth edition edn. 10, 11

Nowak, R. (1995). Genetics: Entering the Postgenome Era. *Science*, **270**, 368–369. 14

Ohyanagi, H., Tanaka, T., Sakai, H., Shigemoto, Y., Yamaguchi, K., Habara, T., Fujii, Y., Antonio, B.A., Nagamura, Y., Imanishi, T., Ikeo, K., Itoh, T., Gojobori, T. & Sasaki, T. (2006). The Rice Annotation Project Database (RAP-DB): hub for Oryza sativa ssp. japonica genome information. *Nucleic acids research*, **34**, D741–4. 34

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D. & Maltsev, N. (1999). The use of gene clusters to infer functional coupling. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 2896–901. 37, 41

Pagani, I., Liolios, K., Jansson, J., Chen, I.M.A., Smirnova, T., Nosrat, B., Markowitz, V.M. & Kyrpides, N.C. (2012). The genomes online database (gold) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Research*, **40**, D571–D579. 16

Park, Y.R., Kim, J., Lee, H.W., Yoon, Y.J. & Kim, J.H. (2011). Gochase-ii: correcting semantic inconsistencies from gene ontology-based annotations for gene products. *BMC Bioinformatics*, **12**, S40. 52

Pearson, H. (2006). Genetics: what is a gene? *Nature*, **441**, 398–401. 10

Pennisi, E. (2000). Ideas Fly at Gene-Finding Jamboree. *Science*, **287**, 2182–2184. 34

Pesquita, C., Faria, D., Bastos, H., Ferreira, A., Falcão, A. & Couto, F. (2008). Metrics for go based protein semantic similarity: a systematic evaluation. *BMC Bioinformatics*, **9**, S4. 28, 29, 54, 68

Pesquita, C., Faria, D., Falcão, A.O., Lord, P. & Couto, F.M. (2009). Semantic Similarity in Biomedical Ontologies. *PLoS Comput Biol*, **5**, e1000443+. 27, 54, 66, 69

Rada, R., Mili, H., Bicknell, E. & Blettner, M. (1989). Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man and Cybernetics*, **19**, 17–30. 25

Rawlings, N.D., Barrett, A.J. & Bateman, A. (2012). Merops: the database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Research*, **40**, 343–350. 5, 103

Raychaudhuri, S. (2003). A literature-based method for assessing the functional coherence of a gene group. *Bioinformatics*, **19**, 396–401. 50

Raychaudhuri, S., Chang, J.T., Sutphin, P.D. & Altman, R.B. (2002). Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature. *Genome research*, **12**, 203–14. 42, 46

Reed, J.L., Famili, I., Thiele, I. & Palsson, B.O. (2006). Towards multidimensional genome annotation. *Nature reviews. Genetics*, **7**, 130–41. 35, 54

Renner, A., Aszódi, A. & Gmbh, N.F. (2000). High-throughput Functional Annotation of Novel Gene Products Using Document Clustering. In *In Proc. Pac. Symp. Biocomputing*, 54–68. 42, 44

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 1*, IJCAI'95, 448–453, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. 3, 26

Rho, K., Kim, B., Jang, Y., Lee, S., Bae, T., Seo, J., Seo, C., Lee, J., Kang, H., Yu, U., Kim, S., Lee, S. & Kim, W. (2011). Garnet – gene set analysis with exploration of annotation relations. *BMC Bioinformatics*, **12**, 1–8. 51

Richards, A.J., Muller, B., Shotwell, M., Cowart, L.A., Rohrer, B. & Lu, X. (2010). Assessing the functional coherence of gene sets with metrics based on the gene ontology graph. *Bioinformatics*, **26**, i79–i87. 48

Riley, M., Abe, T., Arnaud, M.B., Berlyn, M.K.B., Blattner, F.R., Chaudhuri, R.R., Glasner, J.D., Horiuchi, T., Keseler, I.M., Kosuge, T., Mori, H., Perna, N.T., Plunkett, G., Rudd, K.E., Serres, M.H., Thomas, G.H., Thomson, N.R., Wishart, D. & Wanner, B.L. (2006). Escherichia coli K-12: a cooperatively developed annotation snapshot–2005. *Nucleic acids research*, **34**, 1–9. 34

Robinson, P. & Bauer, S. (2011). *Introduction to Bio-Ontologies*. Chapman & Hall/CRC Mathematical & Computational Biology, Taylor & Francis, Boca Raton. 30

Rose, P.W., Bi, C., Bluhm, W.F., Christie, C.H., Dimitropoulos, D., Dutta, S., Green, R.K., Goodsell, D.S., Prlic, A., Quesada, M., Quinn, G.B., Ramos, A.G., Westbrook, J.D., Young, J., Zardecki, C., Berman, H.M. & Bourne, P.E. (2013). The rcsb protein data bank: new resources for research and education. *Nucleic Acids Research*, **41**, D475–D482. 17

Rost, B. (2002). Enzyme function less conserved than anticipated. *Journal of Molecular Biology*, **318**, 595 – 608. 37

Rost, B., Liu, J., Nair, R., Wrzeszczynski, K. & Ofran, Y. (2003). Automatic prediction of protein function. *Cellular and molecular life sciences : CMLS*, **60**, 2637–50. 35

Ruepp, A., Zollner, A., Maier, D., Albermann, K., Hani, J., Mokrejs, M., Tetko, I., Güldener, U., Mannhaupt, G., Münsterkötter, M. & Mewes, H.W. (2004). The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic acids research*, **32**, 5539–45. 22

Ruths, T., Ruths, D. & Nakhleh, L. (2009). Gs2: an efficiently computable measure of go-based similarity of gene sets. *Bioinformatics*, **25**, 1178–1184. 48, 68

Sakata, K., Nagamura, Y., Numa, H., Antonio, B.A., Nagasaki, H., Idonuma, A., Watanabe, W., Shimizu, Y., Horiuchi, I., Matsumoto, T., Sasaki, T. & Higo, K. (2002). RiceGAAS: an automated annotation system and database for rice genome sequence. *Nucleic Acids Research*, **30**, 98–102. 39

Scharf, M., Schneider, R., Casari, G., Bork, P., Valencia, A., Ouzounis, C. & Sander, C. (1994). GeneQuiz: a workbench for sequence analysis. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, **2**, 348–53. 37

Sealfon, R., Hibbs, M., Huttenhower, C., Myers, C. & Troyanskaya, O. (2006). Golem: an interactive graph-based gene-ontology navigation and analysis tool. *BMC Bioinformatics*, **7**, 1–9. 83, 87

Sigrist, C.J.A., Cerutti, L., de Castro, E., Langendijk-Genevaux, P.S., Bulliard, V., Bairoch, A. & Hulo, N. (2010). PROSITE, a protein domain database for functional characterization and annotation. *Nucleic acids research*, **38**, D161–6. 18

Soleti, R., Porro, C. & Martínez, M. (2013). Apoptotic process in cystic fibrosis cells. *Apoptosis*, **18**, 1029–1038. 101

Stam, M.R., Danchin, E.G., Rancurel, C., Coutinho, P.M. & Henrissat, B. (2006). Dividing the large glycoside hydrolase family 13 into subfamilies: towards improved functional annotations of alpha-amylase-related proteins. *Protein Engineering, Design and Selection*, **19**, 555–562. 2, 60

Stein, L. (2001). Genome annotation: from sequence to biology. *Nature reviews. Genetics*, **2**, 493–503. 33, 34

Supek, F., Bošnjak, M., Škunca, N. & Šmuc, T. (2011). Revigo summarizes and visualizes long lists of gene ontology terms. *PLoS ONE*, **6**, e21800. 51

The UniProt Consortium (2013). Update on activities at the universal protein resource (uniprot) in 2013. *Nucleic Acids Research*, **41**, D43–D47. 17

Thomas, P.D., Wood, V., Mungall, C.J., Lewis, S.E., Blake, J.A. & on behalf of the Gene Ontology Consortium (2012). On the use of gene ontology annotations to assess functional similarity among orthologs and paralogs: A short report. *PLoS Comput Biol*, **8**, e1002386. 3

Thornton, J. (2009). Annotations for all by all - the BioSapiens network. *Genome biology*, **10**, 401. 34

Vallenet, D., Labarre, L., Rouy, Z., Barbe, V., Bocs, S., Cruveiller, S., Lajus, A., Pascal, G., Scarpelli, C. & Medigue, C. (2006). MaGe: a microbial genome annotation system supported by synteny results. *Nucleic acids research*, **34**, 53–65. 38

Van Domselaar, G.H., Stothard, P., Shrivastava, S., Cruz, J.A., Guo, A., Dong, X., Lu, P., Szafron, D., Greiner, R. & Wishart, D.S. (2005). BASys: a web server for automated bacterial genome annotation. *Nucleic acids research*, **33**, W455–9. 39

Vidal, A.U., Ettwiller, L. & Birney, E. (2003). Comparative genomics: Genome-wide analysis in metazoan eukaryotes. *Nat Rev Genet*, **4**, 251–262. 39

Vinayagam, A., del Val, C., Schubert, F., Eils, R., Glatting, K.H., Suhai, S. & König, R. (2006). GOPET: a tool for automated predictions of Gene Ontology terms. *BMC bioinformatics*, **7**, 161. 38, 45, 46

Wang, J.Z., Du, Z., Payattakool, R., Yu, P.S. & Chen, C.F. (2007). A new method to measure the semantic similarity of go terms. *Bioinformatics*, **23**, 1274–1281. 48

Webb, E., of Biochemistry, I.U. & Committee, M.B.N. (1992). *Enzyme Nomenclature 1992: Recommendations of the Nciubmb on the Nomenclature and Classification of Enzymes*. Enzyme Nomenclature, Academic Press. 20, 22

Wu, C.H., Huang, H., Arminski, L., Castro-Alvear, J., Chen, Y., Hu, Z.Z., Ledley, R.S., Lewis, K.C., Mewes, H.W., Orcutt, B.C., Suzek, B.E., Tsugita, A., Vinayaka, C.R., Yeh, L.S.L., Zhang, J. & Barker, W.C. (2002). The Protein Information Resource: an integrated public resource of functional annotation of proteins. *Nucleic acids research*, **30**, 35–7. 38

Wu, Z. & Palmer, M. (1994). Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, ACL '94, 133–138, Association for Computational Linguistics, Stroudsburg, PA, USA. 26

Xu, L., Furlotte, N., Lin, Y., Heinrich, K., Berry, M.W., George, E.O. & Homayouni, R. (2011). Functional cohesion of gene sets determined by latent semantic indexing of pubmed abstracts. *PLoS ONE*, **6**, e18851. 50

Yang, Y., Gilbert, D. & Kim, S. (2010). Annotation confidence score for genome annotation: a genome comparison approach. *Bioinformatics*, **26**, 22–29. 50

Yao, Z. & Ruzzo, W.L. (2006). A regression-based K nearest neighbor algorithm for gene function prediction from heterogeneous data. *BMC bioinformatics*, **7 Suppl 1**, S11. 47

Zhang, W., Morris, Q., Chang, R., Shai, O., Bakowski, M., Mitsakakis, N., Mohammad, N., Robinson, M., Zirngibl, R., Somogyi, E., Laurin, N., Eftekharpour, E., Sat, E., Grigull, J., Pan, Q., Peng, W.T., Krogan, N., Greenblatt, J., Fehlings, M., van der Kooy, D., Aubin, J., Bruneau, B., Rossant, J., Blencowe, B., Frey, B. & Hughes, T. (2004). The functional landscape of mouse gene expression. *Journal of biology*, **3**, 21. 37, 41

Zhang, X., Kim, S., Wang, T. & Baral, C. (2006). Joint learning of logic relationships for studying protein function using phylogenetic profiles and the rosetta stone method. *IEEE Transactions on Signal Processing*, **54**, 2427–2435. 37

Zheng, X.H., Lu, F., Wang, Z.Y., Zhong, F., Hoover, J. & Mural, R. (2005). Using shared genomic synteny and shared protein functions to enhance the identification of orthologous gene pairs. *Bioinformatics (Oxford, England)*, **21**, 703–10. 38

Zheng, Y., Roberts, R.J. & Kasif, S. (2002). Genomic functional annotation using co-evolution profiles of gene clusters. *Genome biology*, **3**, 0060.1–9. 41